

TRILL Working Group
INTERNET-DRAFT
Intended Status: Standard Track

Y. Li
D. Eastlake
L. Dunbar
Huawei Technologies
R. Perlman
EMC
M. Umair
IPinfusion
April 17, 2017

Expires: October 19, 2017

TRILL: ARP/ND Optimization
draft-ietf-trill-arp-optimization-08

Abstract

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

Status of this Memo

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

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

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

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

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

Copyright and License Notice

Copyright (c) 2017 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1	Introduction	3
1.1	Terminology	3
2	ARP/ND Optimization Requirement and Solution	4
3	IP/MAC Address Mappings	5
4	Handling ARP/ND/SEND Messages	5
4.1	SEND Considerations	6
4.2	Address Verification	6
4.3	Get Sender's IP/MAC Mapping Information for Non-zero IP	6
4.4	Determine How to Reply to ARP/ND	7
4.5	Determine How to Handle the ARP/ND Response	9
5	Handling RARP (Reverse Address Resolution Protocol) Messages	9
6	Handling of DHCP messages	9
7	Handling of Duplicate IP Addresses	10
8	RBridge ARP/ND Cache Liveness and MAC Mobility	10
9	Security Considerations	11
10	IANA Considerations	11
11	Acknowledgments	11
12	References	12
12.1	Normative References	12
12.2	Informative References	13
	Authors' Addresses	13

1 Introduction

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

1.1 Terminology

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].

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

APPsub-TLV	Application sub-Type-Length-Value [RFC6823]
ARP	Address Resolution Protocol [RFC826]
Campus	A TRILL network consisting of RBridges, links, and possibly bridges bounded by end stations and IP routers [RFC6325]
DAD	Duplicate Address Detection
Data Label	VLAN or FGL
ESADI	End Station Address Distribution Information [RFC7357]
FGL	Fine-Grained Label [RFC7172]
IA	Interface Addresses, a TRILL APPsub-TLV [RFC7961]
IP	Internet Protocol, both IPv4 and IPv6
MAC	Media Access Control [RFC7042]
ND	Neighbor Discovery [RFC4861]
RBridge	A contraction of "Routing Bridge". A device implementing the TRILL protocol.
SEND	secure neighbor discovery [RFC3971]

TRILL Transparent Interconnection of Lots of Links or
Tunneled Routing in the Link Layer [RFC6325] [RFC7780]

2 ARP/ND Optimization Requirement and Solution

IP address resolution can create significant issues in data centers due to flooded packets as discussed in [RFC6820]. Such flooding can be avoided by a proxy ARP/ND function on edge RBridges as described in this document.

To support such ARP/ND optimization, edge RBridges need to know end-station's IP to MAC mapping through manual configuration (management), through control plane mechanisms such as directories [DirMech], or through Data plane learning by snooping of messages such as ARP/ND (including DHCP or gratuitous ARP_messages).

When all the end-stations IP/MAC address mapping is known to edge RBridges or provisioned through management or learnt via control plane on the edge RBridges, it should be possible to completely suppress flooding of ARP/ND messages in a TRILL Campus, When all end-station MAC addresses are similarly known, it should be possible to suppress unknown unicast flooding by dropping any unknown unicast received at an edge RBridge.

An ARP/ND optimization mechanism should include provisions for an edge RBridge to issue an ARP/ND request to an attached end station to confirm or update information and should allow an end station to detect duplicate IP addresses.

TRILL already provides an option to disable data-plane learning from the source MAC address of end-station frames on a per port basis (see Section 5.3 of [RFC6325]).

Most of the end station hosts either send DHCP messages requesting an IP Address or send out gratuitous ARP or RARP requests to announce themselves to the network right after they come online. Thus the local edge RBridge will immediately have the opportunity to snoop and learn their MAC and IP addresses and distribute this information to other edge RBridges through the TRILL control plane ESADI [RFC7357] protocol. Once remote RBridges received this information via the control plane they should add IP to MAC mapping information to their ARP/ND cache along with the nickname and data label of the address information. Therefore, most active IP hosts in TRILL network can be learned by the edge RBridges either through local learning or control-plane-based remote learning. As a result, ARP suppression can vastly reduce the network flooding caused by host ARP learning behavior.

3 IP/MAC Address Mappings

By default, an RBridge [RFC6325] [RFC7172] learns MAC Address and Data Label (VLAN or FGL) to egress nickname mapping information from TRILL data frames it receives. No IP address information is learned directly from the TRILL data frame. The Interface Addresses (IA) APPsub-TLV [RFC7961] enhances the TRILL base protocol by allowing IP and MAC address mappings to be distributed in the control plane by any RBridge. This APPsub-TLV appears inside the TRILL GENINFO TLV in ESADI [RFC7357] but the value data structure it specifies may also occur in other application contexts. Edge RBridge Directory Assist Mechanisms [DirMech] makes use of this APPsub-TLV for its push model and uses the value data structure it specifies in its pull model.

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

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

4 Handling ARP/ND/SEND Messages

A native frame that is an ARP [RFC826] message is detected by its Ethertype of 0x0806. A native frame that is an ND [RFC4861] is detected by being one of five different ICMPv6 packet types. ARP/ND is commonly used on a link to (1) query for the MAC address corresponding to an IPv4 or IPv6 address, (2) test if an IPv4/IPv6 address is already in use, or (3) to announce the new or updated info on any of IPv4/IPv6 address, MAC address, and/or point of attachment.

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

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

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

4.1 SEND Considerations

SEND (Secure Neighbor Discovery [RFC3971] is a method of securing ND that addresses the threats discussed in [RFC3756]. Typical TRILL campuses are inside data centers, Internet exchange points, or carrier facilities. These are generally controlled and protected environments where these threats are of less concern. Nevertheless, SEND provides an additional layer of protection.

Secure SEND messages require knowledge of cryptographic keys. Methods of communicating such keys to RBridges for use in SEND are beyond the scope of this document. Thus, using the optimizations in this document, RBridges do not attempt to construct SEND messages and are generally transparent to them. RBridges only construct ARP, RARP, or insecure ND messages, as appropriate. Nevertheless, RBridges implementing ARP/ND optimization SHOULD snoop on SEND messages to extract addressing information that would be present if the message had been sent as an insecure ND message.

4.2 Address Verification

RBridges may use ARP/ND to probe directly attached or remote end stations for address or liveness verification. This is typically most appropriate in less managed and/or higher mobility environments. In strongly managed environments, such as a typical data center, where a central orchestration/directory system has complete addressing knowledge [RFC7067], optimized ARP/ND responses can use that knowledge. In such cases, there is little reason for verification except for debugging operational problems or the like.

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

- o If the sender's IP is not present in the ingress RBridge's ARP/ND

cache, populate the information of sender's IP/MAC in its ARP/ND cache table. The ingress RBridge correlates its nickname and that IP/MAC mapping information. Such triplet of {IP address, MAC address, ingress nickname} information is saved locally and can be distributed to other RBridges as explain later.

o Else if the sender's IP has been saved before but with a different MAC address mapped or a different ingress nickname associated with the same pair of IP/MAC, the RBridge SHOULD verify if a duplicate IP address has already been in use or an end station has changed its attaching RBridge. The RBridge may use different strategies to do so. For example, the RBridge might ask an authoritative entity like directory servers or it might encapsulate and unicast the ARP/ND message to the location where it believes the address is in use. RBridge SHOULD update the saved triplet of {IP address, MAC address, ingress nickname} based on the verification. An RBridge might not verify an IP address if the network manager's policy is to have the network behave, for each Data Label, as if it were a single link and just believe an ARP/ND it receives.

The ingress RBridge MAY use the IA APPsub-TLV [RFC7961] with the Local flag set in ESADI [RFC7357] to distribute any new or updated triplet of {IP address, MAC address, ingress nickname} information obtained in this step. If a push directory server is used, such information can be distributed as per [DirMech].

4.4 Determine How to Reply to ARP/ND

The options for an edge RBridge to handle a native ARP/ND are given below. For generic ARP/ND request seeking the MAC address corresponding to an IP address, if the edge RBridge knows the IP address and corresponding MAC, behavior is as in item (a), otherwise behavior is as in item (b). Behavior for gratuitous ARP and ND Unsolicited Neighbor Advertisements [RFC4861] is given in item (c). And item (d) covers handling of Address Probe ARP Query.

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

a) If the message is a generic ARP/ND request and the ingress RBridge knows the target's IP address and associated MAC address, the ingress RBridge MUST take one or a combination of the actions below. In the case of secure neighbor discovery (SEND) [RFC3971], cryptography would prevent local reply by the ingress RBridge, since the RBridge would not be able to sign the response with the target's private key, and only action a.2 or a.5 is valid.

a.1. Send an ARP/ND response directly to the querier, using the target's MAC address present in the ingress RBridge's ARP/ND cache table. Because the edge RBridge might not have an IPv6 address, the source IP address for such an ND response MUST be that of the target end station.

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

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

a.4. Reply to the querier based on directory information [DirMech] such as information obtained from a pull directory server or directory information that the ingress RBridge has requested to be pushed to it.

a.5. Flood the /ND/SEND request as per [RFC6325].

(b) If the message is a generic ARP/ND/SEND request and the ingress RBridge does not know target's IP address, the ingress RBridge MUST take one of the following actions. In the case of secure neighbor discovery (SEND) [RFC3971], cryptography would prevent local reply by the ingress RBridge, since the RBridge would not be able to sign the response with the target's private key therefore only action b.1 is valid.

b.1. Flood the ARP/ND/SEND message as per [RFC6325].

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

b.3. Drop the message if the directory mechanism is used and you know there should be no response (query based on a non-existent IP address for example).

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

The RBridge MAY use an IA APPsub-TLV [RFC7961] with the Local flag

set to distribute the sender's MAC and IP mapping information. When one or more directory servers are deployed and complete Push Directory information is used by all the RBridges in the Data Label, a gratuitous ARP or unsolicited NA SHOULD be discarded rather than ingressed. Otherwise, they are either ingressed and flooded as per [RFC6325] or discarded depending on local policy.

(d) If the message is a Address Probe ARP Query [RFC5227] which can be identified by the sender's protocol (IPv4) address field being zero and the target's protocol address field being the IPv4 address to be tested or a Neighbor Solicitation for DAD (Duplicate Address Detection) which has the unspecified source address [RFC4862]: it SHOULD be handled as the generic ARP message as in (a) or (b) above.

4.5 Determine How to Handle the ARP/ND Response

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

5 Handling RARP (Reverse Address Resolution Protocol) Messages

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

6 Handling of DHCP messages

When a newly connected end-station exchanges messages with a DHCP [RFC2131] server an edge RBridge should snoop them (mainly the DHCPACK message) and store IP/MAC mapping information in its ARP/ND

cache and should also send the information out through the TRILL control plane using ESADI.

7 Handling of Duplicate IP Addresses

Duplicate IP addresses within a Data Label can occur due to an attacker sending fake ARP/ND messages or due to human/configuration errors. If complete directory information is available, then by definition the IP location information in the directory is correct. Any appearance of an IP address in a different place (different edge RBridge or port) from other sources is not correct.

Without complete directory information, the ARP/ND optimization function should support duplicate IP detection. This is critical in a Data Center to stop an attacker from using ARP/ND spoofing to divert traffic from its intended destination.

Duplicate IP addresses can be detected when an existing active IP1/MAC1 mapping gets modified. Also an edge RBridge may send a query to the former owner of IP called a DAD-query (Duplicate Address Detection query). A DAD-query is a unicast ARP/ND message with sender IP 0.0.0.0 in case of ARP (or a configurable per RBridge IP address called the DAD-Query source IP) and an IPv6 Link Local Address in case of ND with source MAC set to the DAD-querier RBridge's MAC. If the querying RBridge does not receive an answer within a given time, the new IP entry will be confirmed and activated in its ARP/ND cache.

In the case where the former owner replies, a Duplicate Address has been detected. In this case the querying RBridge SHOULD log the duplicate so that the network administrator can take appropriate action.

8 RBridge ARP/ND Cache Liveness and MAC Mobility

A maintenance procedure is needed for RBridge ARP/ND caching to ensure IP end-stations connected to ingress RBridges are still active.

Some links provide a physical layer indication of link liveness. A dynamic proxy-ARP/ND entry (one learned from data plane observation) MUST be removed from the table if the link over which it was learned fails.

Similarly a dynamic proxy-ARP/ND entry SHOULD be flushed out of the table if the IP/MAC mapping has not been refreshed within a given age-time. The entry is refreshed if an ARP or ND message is received for the same IP/MAC mapping entry from any location. The IP/MAC

mapping information ageing timer is configurable per RBridge and defaults to 3/4 of the MAC address learning Ageing Timer [RFC6325].

For example end-Station "A" is connected to edge-RBridge1 (RB1) and has been learnt as local entry on RB1. If end-Station "A" moves to some other location (MAC/VM Mobility) and gets connected to edge-RBridge2 (RB2), after learning on RB2's access port, RB2 advertises this entry through the TRILL control-plane and it gets learnt on RB1 as a remote entry. The old entry on RB1 SHOULD get replaced and all other edge-RBridges with end-station service enabled for that data-label should update the entry to show reachability from RB2 instead of RB1.

If an ARP/ND entry in the cache is not refreshed, then the RBridge connected to that end-station MAY send periodic refresh messages (ARP/ND "probes") to that end-station, so that the entries can be refreshed before they age out. The end-station would reply to the ARP/ND probe and the reply resets the corresponding entry age-timer.

9 Security Considerations

Unless Secure ND (SEND [RFC3971]) is used, ARP and ND messages can be easily forged. Therefore the learning of MAC/IP addresses by RBridges from ARP/ND should not be considered as reliable. See Section 4.1 for SEND Considerations.

An RBridge can use the confidence level in IA APPsub-TLV information received via ESADI or pull directory retrievals to determine the reliability of MAC/IP address mapping. ESADI information can be secured as provided in [RFC7357] and pull directory information can be secured as provided in [DirMech]. The implementation decides if an RBridge will distribute the IP and MAC address mappings received from local native ARP/ND messages to other RBridges in the same Data Label, if it distributes them, and with what confidence level it does so.

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

10 IANA Considerations

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

11 Acknowledgments

The authors would like to thank Igor Gashinsky and Sue Hares for their contributions.

12 References

12.1 Normative References

- [RFC826] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.
- [RFC903] Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): 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.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, September 2014.
- [RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014.
- [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, February 2016.

- [RFC7961] Eastlake 3rd, D. and L. Yizhou, "Transparent Interconnection of Lots of Links (TRILL): Interface Addresses APPsub-TLV", RFC 7961, August 2016.
- [DirMech] Dunbar, L., Eastlake 3rd, D., Perlman, R., I. Gashinsky. and Li Y., "TRILL: Edge Directory Assist Mechanisms", draft-ietf-trill-directory-assist-mechanisms, work in progress.

12.2 Informative References

- [RFC3756] Nikander, P., Ed., Kempf, J., and E. Nordmark, "IPv6 Neighbor Discovery (ND) Trust Models and Threats", RFC 3756, May 2004.
- [RFC3971] Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.
- [RFC5227] Cheshire, S., "IPv4 Address Conflict Detection", RFC 5227, July 2008.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, January 2013.
- [RFC6823] Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, December 2012.
- [RFC7042] Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, October 2013.
- [RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, November 2013.

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56625375
EMail: liyizhou@huawei.com

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

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

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

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

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

EMail: Radia@alum.mit.edu

Mohammed Umair
IPinfusion

Email: mohammed.umair2@gmail.com

TRILL Working Group
INTERNET-DRAFT
Intended Status: Standard Track

Y. Li
D. Eastlake
L. Dunbar
Huawei Technologies
R. Perlman
EMC
M. Umair
Cisco
October 9, 2017

Expires: April 12, 2017

TRILL: ARP/ND Optimization
draft-ietf-trill-arp-optimization-09

Abstract

This document describes mechanisms to optimize the ARP (Address Resolution Protocol) and ND (Neighbor Discovery) traffic in a TRILL campus. TRILL switches maintain a cache of IP/MAC address/Data Label bindings that are learned from ARP/ND requests and responses that pass through them. In many cases, this cache allows an edge RBridge to avoid flooding an ARP/ND request by either responding to it directly or by encapsulating it and unicasting it. Such optimization reduces packet flooding over a TRILL campus.

Status of this Memo

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

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

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

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

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

Copyright and License Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1	Introduction	3
1.1	Terminology	3
2	ARP/ND Optimization Requirement and Solution	4
3	IP/MAC Address Mappings	5
4	Handling ARP/ND/SEND Messages	5
4.1	SEND Considerations	6
4.2	Address Verification	7
4.3	Extracting Local End Station IP/MAC Mapping Information	7
4.4	Determine How to Reply to ARP/ND	8
4.5	Determine How to Handle the ARP/ND Response	10
5	Handling of RARP (Reverse Address Resolution Protocol) Messages	10
6	Handling of DHCP messages	10
7	Handling of Duplicate IP Addresses	10
8	RBridge ARP/ND Cache Liveness and MAC Mobility	11
9	Security Considerations	12
9.1	Data Plane Based Considerations	12
9.2	Directory Based Considerations	13
9.3	Use of the Confidence Level Feature	13
10	IANA Considerations	13
11	Acknowledgments	14
12	References	14
12.1	Normative References	14
12.2	Informative References	15
	Authors' Addresses	16

1 Introduction

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

1.1 Terminology

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].

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

APPsub-TLV	Application sub-Type-Length-Value [RFC6823]
ARP	Address Resolution Protocol [RFC826]
Campus	A TRILL network consisting of RBridges, links, and possibly bridges bounded by end stations and IP routers [RFC6325]
DAD	Duplicate Address Detection
Data Label	VLAN or FGL
DHCP	In this document refers to both DHCP for IPv4 [RFC2131] and DHCPv6 [RFC3315]
ESADI	End Station Address Distribution Information [RFC7357]
FGL	Fine-Grained Label [RFC7172]
IA	Interface Addresses, a TRILL APPsub-TLV [RFC7961]
IP	Internet Protocol, both IPv4 and IPv6
MAC	Media Access Control [RFC7042]
ND	Neighbor Discovery [RFC4861]
RBridge	A contraction of "Routing Bridge". A device

implementing the TRILL protocol.

SEND secure neighbor discovery [RFC3971]

TRILL Transparent Interconnection of Lots of Links or
Tunneled Routing in the Link Layer [RFC6325] [RFC7780]

2 ARP/ND Optimization Requirement and Solution

IP address resolution can create significant issues in data centers due to flooded packets as discussed in [RFC6820]. Such flooding can be avoided by a proxy ARP/ND function on edge RBridges as described in this document. This section is a general discussion of this problem and is not intended to be normative.

To support such ARP/ND optimization, edge RBridges need to know end-station's IP to MAC mapping through manual configuration (management), through control plane mechanisms such as directories [RFC8171], or through Data plane learning by snooping of messages such as ARP/ND (including DHCP or gratuitous ARP messages).

When all the end-stations IP/MAC address mapping is known to edge RBridges or provisioned through management or learnt via control plane on the edge RBridges, it should be possible to completely suppress flooding of ARP/ND messages in a TRILL Campus. When all end-station MAC addresses are similarly known, it should be possible to suppress unknown unicast flooding by dropping any unknown unicast received at an edge RBridge.

An ARP/ND optimization mechanism should include provisions for an edge RBridge to issue an ARP/ND request to an attached end station to confirm or update information and should allow an end station to detect detect duplication of its IP address.

Most of the end station hosts either send DHCP messages requesting an IP Address or send out gratuitous ARP or RARP requests to announce themselves to the network right after they come online. Thus the local edge RBridge will immediately have the opportunity to snoop and learn their MAC and IP addresses and distribute this information to other edge RBridges through the TRILL control plane ESADI [RFC7357] protocol. Once remote RBridges received this information via the control plane they should add IP to MAC mapping information to their ARP/ND cache along with the nickname and data label of the address information. Therefore, most active IP hosts in TRILL network can be learned by the edge RBridges either through local learning or control-plane-based remote learning. As a result, ARP suppression can vastly reduce the network flooding caused by host ARP learning behavior.

When complete directory information is available, the default data plane learning of MAC addresses of end station is not only unnecessary but could be harmful if there is learning based on frames with forged source addresses. Such data plane learning can be suppressed because TRILL already provides an option to disable data-plane learning from the source MAC address of end-station frames (see Section 5.3 of [RFC6325]).

3 IP/MAC Address Mappings

By default, an RBridge [RFC6325] [RFC7172] learns MAC Address and Data Label (VLAN or FGL) to egress nickname mapping information from TRILL data frames it receives. No IP address information is learned directly from the TRILL data frame. The Interface Addresses (IA) APPsub-TLV [RFC7961] enhances the TRILL base protocol by allowing IP and MAC address mappings to be distributed in the control plane by any RBridge. This APPsub-TLV appears inside the TRILL GENINFO TLV in ESADI [RFC7357] but the value data structure it specifies may also occur in other application contexts. Edge RBridge Directory Assist Mechanisms [RFC8171] makes use of this APPsub-TLV for its push model and uses the value data structure it specifies in its pull model.

An RBridge can easily know the IP/MAC address mappings of the local end stations that it is attached to it via its access ports by receiving ARP [RFC826] or ND [RFC4861] messages. If the edge RBridge has extracted the sender's IP/MAC address pair from the received data frame (either ARP or ND), it may save the information and then use the IA APPsub-TLV to link the IP and MAC addresses and distribute it to other RBridges through ESADI. Then the relevant remote RBridges (normally those interested in the same Data Label as the original ARP/ND messages) also receive and save such mapping information. There are others ways that RBridges save IP/MAC address mappings in advance, e.g. import from management system and distribution by directory servers [RFC8171].

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

4 Handling ARP/ND/SEND Messages

A native frame that is an ARP [RFC826] message is detected by its Ethertype of 0x0806. A native frame that is an ND [RFC4861] is detected by being one of five different ICMPv6 packet types. ARP/ND is commonly used on a link to (1) query for the MAC address corresponding to an IPv4 or IPv6 address, (2) test if an IPv4/IPv6 address is already in use, or (3) to announce the new or updated info on any of IPv4/IPv6 address, MAC address, and/or point of attachment.

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

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

When an ingress RBridge receives an ARP/ND/SEND message, it can perform the steps described within the sub-sections below. In particular, Section 4.4 describes the options for such an ingress handling an ARP/ND message and, in the cases where it forwards the message, Section 4.5 describes how to handle any response that may be returned due to the forwarded message.

Section 4.3 describes the extraction of address information by an RBridge from ARP/ND messages it handles. Under some circumstances, this extraction may prompt verification with an end station. Section 4.2 describes an optional use of ARP/ND messages originated by RBridges to verify addresses or liveness.

As described in Section 4.1, SEND messages are not optimized by the mechanisms specified in this document but are snooped on.

4.1 SEND Considerations

SEND (Secure Neighbor Discovery [RFC3971] is a method of securing ND that addresses the threats discussed in [RFC3756]. Typical TRILL campuses are inside data centers, Internet exchange points, or carrier facilities. These are generally controlled and protected environments where these threats are of less concern. Nevertheless, SEND provides an additional layer of protection.

Secure SEND messages require knowledge of cryptographic keys. Methods of communicating such keys to RBridges for use in SEND are beyond the

scope of this document. Thus, using the optimizations in this document, RBridges do not attempt to construct SEND messages and are generally transparent to them. RBridges only construct ARP, RARP, or insecure ND messages, as appropriate. Nevertheless, RBridges implementing ARP/ND optimization SHOULD snoop on SEND messages to extract the addressing information that would be present if the SEND had been sent as an insecure ND message and is still present in the SEND message.

4.2 Address Verification

RBridges may use ARP/ND to probe directly attached or remote end stations for address or liveness verification. This is typically most appropriate in less managed and/or higher mobility environments. In strongly managed environments, such as a typical data center, where a central orchestration/directory system has complete addressing knowledge [RFC7067], optimized ARP/ND responses can use that knowledge. In such cases, there is little reason for verification except for debugging operational problems or the like.

4.3 Extracting Local End Station IP/MAC Mapping Information

Edge RBridges extract and use information about the correspondence between local end station IP and MAC addresses from the ARP/ND messages those end stations send as described below. An apparent zero source IP address means that the end station is probing for duplicate IP addresses and messages with such a zero source IP address are not used for the extraction of IP/MAC address mapping information.

- o If the sender's IP is not present in the ingress RBridge's ARP/ND cache, populate the information of sender's IP/MAC in its ARP/ND cache table. The ingress RBridge correlates its nickname and that IP/MAC mapping information. Such triplet of {IP address, MAC address, ingress nickname} information is saved locally and can be distributed to other RBridges as explain later.

- o Else if the sender's IP has been saved before but with a different MAC address mapped or a different ingress nickname associated with the same pair of IP/MAC, the RBridge SHOULD verify if a duplicate IP address has already been in use or an end station has changed its attaching RBridge. The RBridge may use different strategies to do so. For example, the RBridge might ask an authoritative entity like directory servers or it might encapsulate and unicast the ARP/ND message to the location where it believes the address is in use (Section 4.2). RBridge SHOULD update the saved

triplet of {IP address, MAC address, ingress nickname} based on the verification results. An RBridge might not verify an IP address if the network manager's policy is to have the network behave, for each Data Label, as if it were a single link and just believe an ARP/ND it receives.

The ingress RBridge MAY use the IA APPsub-TLV [RFC7961] with the Local flag set in ESADI [RFC7357] to distribute any new or updated triplet of {IP address, MAC address, ingress nickname} information obtained. If a push directory server is used, such information can be distributed as specified in [RFC8171].

4.4 Determine How to Reply to ARP/ND

The options for an edge RBridge to handle a native ARP/ND are given below. For generic ARP/ND request seeking the MAC address corresponding to an IP address, if the edge RBridge knows the IP address and corresponding MAC, behavior is as in item (a), otherwise behavior is as in item (b). Behavior for gratuitous ARP and ND Unsolicited Neighbor Advertisements [RFC4861] is given in item (c). And item (d) covers handling of Address Probe ARP Query. Within each lettered item, it is an implementation decision which numbered strategy to use for any particular ARP/ND query falling under that item.

a) If the message is a generic ARP/ND request and the ingress RBridge knows the target's IP address and associated MAC address, the ingress RBridge MUST take one or a combination of the actions below. In the case of secure neighbor discovery (SEND) [RFC3971], cryptography would prevent local reply by the ingress RBridge, since the RBridge would not be able to sign the response with the target's private key, and only action a.2 or a.5 is valid.

a.1. Send an ARP/ND response directly to the querier, using the target's MAC address present in the ingress RBridge's ARP/ND cache table. Because the edge RBridge might not have an IPv6 address, the source IP address for such an ND response MUST be that of the target end station.

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

a.3. Block ARP/ND requests that occur for some time after a request to the same target has been launched, and then respond to the

querier when the response to the recently-launched query to that target is received.

a.4 Reply to the querier based on directory information [RFC8171] such as information obtained from a pull directory server or directory information that the ingress RBridge has requested to be pushed to it.

a.5. Flood the ARP/ND/SEND request as per [RFC6325].

(b) If the message is a generic ARP/ND/SEND request and the ingress RBridge does not know target's IP address, the ingress RBridge MUST take one of the following actions. In the case of secure neighbor discovery (SEND) [RFC3971], cryptography would prevent local reply by the ingress RBridge, since the RBridge would not be able to sign the response with the target's private key therefore only action b.1 is valid.

b.1. Flood the ARP/ND/SEND message as per [RFC6325].

b.2. Use directory server to pull the information [RFC8171] and reply to the querier.

b.3. Drop the message if there should be no response because the directory server gives authoritative information that the address being queried is non-existent.

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

The RBridge MAY use an IA APPsub-TLV [RFC7961] with the Local flag set to distribute the sender's MAC and IP mapping information. When one or more directory servers are deployed and complete Push Directory information is used by all the RBridges in the Data Label, a gratuitous ARP or unsolicited NA SHOULD be discarded rather than ingressed. Otherwise, they are either ingressed and flooded as per [RFC6325] or discarded depending on local policy.

(d) If the message is a Address Probe ARP Query [RFC5227] which can be identified by the sender's protocol (IPv4) address field being zero and the target's protocol address field being the IPv4 address to be tested or a Neighbor Solicitation for DAD (Duplicate Address Detection) which has the unspecified source address [RFC4862]: it SHOULD be handled as the generic ARP message as in (a) or (b) above.

4.5 Determine How to Handle the ARP/ND Response

If the ingress RBridge R1 decides to unicast the ARP/ND request to the target's egress RBridge R2 as discussed in subsection 3.2 item a.2 or to flood the request as per [RFC6325] and item a.5, then R2 decapsulates the query, and initiates an ARP/ND query on the target's link. If and when the target responds, R2 encapsulates and unicasts the response to R1, which decapsulates the response and sends it to the querier. R2 SHOULD initiate a link state update to inform all the other RBridges of the target's location, layer 3 address, and layer 2 address, in addition to forwarding the reply to the querier. The update uses an IA APPsub-TLV [IA-draft] (so the layer 3 and layer 2 addresses can be linked) with the Local flag set in ESADI [RFC7357] or as per [RFC8171] if push directory server is in use.

5 Handling of RARP (Reverse Address Resolution Protocol) Messages

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

6 Handling of DHCP messages

When a newly connected end-station exchanges messages with a DHCP [RFC3315][RFC2131] server an edge RBridge should snoop them (mainly the DHCPack message) and store IP/MAC mapping information in its ARP/ND cache and should also send the information out through the TRILL control plane using ESADI.

7 Handling of Duplicate IP Addresses

Duplicate IP addresses within a Data Label can occur due to an attacker sending fake ARP/ND messages or due to human/configuration errors. If complete directory information is available, then by definition the IP location information in the directory is correct. Any appearance of an IP address in a different place (different edge RBridge or port) from other sources is not correct.

Without complete directory information, the ARP/ND optimization function should support duplicate IP detection. This is critical in a Data Center to stop an attacker from using ARP/ND spoofing to divert traffic from its intended destination.

Duplicate IP addresses can be detected when an existing active IP/MAC mapping gets modified. Also an edge RBridge may send a query called a DAD-query (Duplicate Address Detection query) asking about the IP address in question to the former owner of that IP address by using the MAC address formerly associated with that IP address. A DAD-query is a unicast ARP/ND message with sender IP 0.0.0.0 in case of ARP (or a configurable IP address per RBridge called the DAD-Query source IP) and an IPv6 Link Local Address in case of ND with source MAC set to the DAD-querier RBridge's MAC. If the querying RBridge does not receive an answer within a given time, it may be a case of mobility and in any case the new IP entry will be confirmed and activated in its ARP/ND cache.

In the case where the former owner replies, a Duplicate Address has been detected. In this case the querying RBridge SHOULD log the duplicate so that the network administrator can take appropriate action.

It is an implementation choice how to respond to a query for an address that is duplicated in the network when authoritative information is not available from a directory or configuration. Typically the information most recently snooped is returned.

8 RBridge ARP/ND Cache Liveness and MAC Mobility

A maintenance procedure is needed for RBridge ARP/ND caching to ensure IP end-stations connected to ingress RBridges are still active.

Some links provide a physical layer indication of link liveness. A dynamic proxy-ARP/ND entry (one learned from data plane observation) MUST be removed from the table if the link over which it was learned fails.

Similarly a dynamic proxy-ARP/ND entry SHOULD be flushed out of the table if the IP/MAC mapping has not been refreshed within a given age-time. The entry is refreshed if an ARP or ND message is received for the same IP/MAC mapping entry from any location. The IP/MAC mapping information ageing timer is configurable per RBridge and defaults to 3/4 of the MAC address learning Ageing Timer [RFC6325].

For example end-Station "A" is connected to edge-RBridge1 (RB1) and has been learnt as local entry on RB1. If end-Station "A" moves to some other location (MAC/VM Mobility) and gets connected to edge-RBridge2 (RB2), after learning on RB2's access port, RB2 advertise this entry through the TRILL control-plane and it gets learnt on RB1 as a remote entry. The old entry on RB1 SHOULD get replaced and all other edge-RBridges with end-station service enabled for that data-

label should update the entry to show reachability from RB2 instead of RB1.

If an ARP/ND entry in the cache is not refreshed, then the RBridge connected to that end-station MAY send periodic refresh messages (ARP/ND "probes") to that end-station, so that the entries can be refreshed before they age out. The end-station would reply to the ARP/ND probe and the reply resets the corresponding entry age-timer.

9 Security Considerations

There are generally two modes of learning the address information that is the basis of ARP/ND optimization: data plane mode and directory mode. The data plane mode is the traditional bridge address learning [802.1Q] that is also implemented in TRILL switches [RFC6325] and is discussed in Section 9.1. The directory mode uses data obtained from a directory [RFC8171] and is discussed in Section 9.2. The TRILL confidence level feature, which can help arbitrate between conflicting address information, is discussed in Section 9.3.

RBridges should rate limit of ARP/ND queries injected into the TRILL campus to limit some potential denial of service attacks.

9.1 Data Plane Based Considerations

Generally speaking, when ARP/ND optimization is operating in the data plane mode, the information learned by RBridges is the same as that which is learned by end stations. Thus the answers generated by RBridges to the query messages being optimized are generally those that would be generated by end stations in the absence of optimization and the security considerations are those of the underlying ARP/ND protocols.

RBridges that snoop on DHCPack messages respond to ARP/ND messages in essentially the same way that the end stations sending those DHCPack messages would. Thus, for Security Considerations of ARP/ND optimization for DHCP messages that may be snooped, see the Security Considerations sections of [RFC3315] and [RFC2131].

Unless Secure ND (SEND [RFC3971]) is used, ARP and ND messages can be easily forged. Therefore the learning of MAC/IP addresses by RBridges from ARP/ND is hackable but is what is available for data plane learning without SEND. See Section 4.1 for SEND Considerations.

Since end stations communicate with edge RBridges using Ethernet, some security improvement could be obtained by the use of [802.1AE] between end stations and edge RBridges. Such link security would

impose requirements on edge stations, while TRILL is generally designed to operate with unmodified, TRILL-ignorant end stations, and is beyond the scope of this document

ARP/ND address mapping information learned locally at an RBridge can be distributed to other RBridges using the TRILL ESADI protocol that can be secured as specified in [RFC7357]. (ESADI is also used for push directories with flags in the data indicating whether data come from a directory or from data plane learning, as well as a confidence level (see Section 9.3).)

9.2 Directory Based Considerations

ARP/ND optimization can be based on directory information [RFC8171]. If the directory information is known to be trustworthy and complete, then trustworthy responses to ARP/ND queries can be entirely based on this information. This bounds the damage that forged ARP/ND messages can do to the local link between end stations and edge RBridges. (In TRILL, such a "link" can be a bridged LAN.)

Of course, there can also be incomplete and/or un-reliable directory address mapping data. The network administrator can configure their TRILL campus to use such directory data in place of data plane learned data. Alternatively, such directory data can be used along with data plane learned arbitrated by confidence level as discussed in Section 9.3.

9.3 Use of the Confidence Level Feature

An RBridge can use the confidence level in IA APPsub-TLV information received via ESADI or pull directory retrievals to determine the configured relative reliability of MAC/IP address mapping information from those sources and from locally learned address information. ESADI / push directory information can be secured as provided in [RFC7357] and pull directory information can be secured as provided in [RFC8171]. The implementation decides if an RBridge will distribute the IP and MAC address mappings received from local native ARP/ND messages to other RBridges in the same Data Label, and with what confidence level it does so. Thus the implementer can, to some extent, cause sources that they know are more reliable to dominate those they know to be less reliable. How the implementer determines this is beyond the scope of this document.

10 IANA Considerations

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

11 Acknowledgments

The authors would like to thank Igor Gashinsky and Sue Hares for their contributions.

12 References

12.1 Normative References

- [RFC826] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.
- [RFC903] Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.
- [RFC3315] Droms, R., Ed., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): 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.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, September 2014.
- [RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O.

Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014.

- [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, February 2016.
- [RFC7961] Eastlake 3rd, D. and L. Yizhou, "Transparent Interconnection of Lots of Links (TRILL): Interface Addresses APPsub-TLV", RFC 7961, August 2016.
- [RFC8171] Eastlake 3rd, D., Dunbar, L., Perlman, R., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Edge Directory Assistance Mechanisms", RFC 8171, June 2017.

12.2 Informative References

- [802.1AE] IEEE Std 802.1AE-2006, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security, 18 August 2006.
- [802.1Q] IEE Std 802.1Q-2014, IEEE Standard for Local and metropolitan area networks / Bridges and Bridged Networks, 3 November 2014.
- [RFC3756] Nikander, P., Ed., Kempf, J., and E. Nordmark, "IPv6 Neighbor Discovery (ND) Trust Models and Threats", RFC 3756, May 2004.
- [RFC3971] Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.
- [RFC5227] Cheshire, S., "IPv4 Address Conflict Detection", RFC 5227, July 2008.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, January 2013.
- [RFC6823] Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, December 2012.

[RFC7042] Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, October 2013.

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

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56625375
EMail: liyizhou@huawei.com

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

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

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

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

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

EMail: Radia@alum.mit.edu

Mohammed Umair
Cisco
Cessna Business Park, Kadubeesanahalli Village, Hobli,
Sarjapur, Varthur Main Road, Marathahalli,

Bengaluru, Karnataka 560087 India

Email: mohammed.umair2@gmail.com

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

Margaret Cullen
Painless Security
Donald Eastlake
Mingui Zhang
Dacheng Zhang
Huawei
May 31, 2017

Expires: November 30, 2017

TRILL (Transparent Interconnection of Lots of Links) over IP
<draft-ietf-trill-over-ip-10.txt>

Abstract

The TRILL (Transparent Interconnection of Lots of Links) 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 specifies transmission of encapsulated TRILL data and TRILL IS-IS over IP (v4 or v6). so as to use an IP network as a TRILL link in a unified TRILL campus. This document 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 authors or the TRILL Working Group 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.....	7
3.3.1 Security Requirements.....	7
3.3.2 Multicast Handling.....	8
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.....	17
5.5 VXLAN Encapsulation.....	18
5.6 TCP Encapsulation.....	18
5.7 Other Encapsulations.....	19
6. Handling Multicast.....	20
7. Use of IPsec and IKEv2.....	21
7.1 Keying.....	21
7.1.1 Pairwise Keying.....	21
7.1.2 Group Keying.....	22
7.2 Mandatory-to-Implement Algorithms.....	22
8. Transport Considerations.....	23
8.1 Congestion Considerations.....	23
8.1.1 Within a TMCE.....	24
8.1.2 In Other Environments.....	24
8.2 Recursive Ingress.....	24
8.3 Fat Flows.....	25
8.4 MTU Considerations.....	26
8.5 Middlebox Considerations.....	27

Table of Contents (continued)

9. TRILL over IP Port Configuration.....	28
9.1 Per IP Port Configuration.....	28
9.2 Additional per IP Address Configuration.....	28
9.2.1 Native Multicast Configuration.....	29
9.2.2 Serial Unicast Configuration.....	29
9.2.3 Encapsulation Specific Configuration.....	29
9.2.3.1 UDP and TCP Source Port.....	29
9.2.3.2 VXLAN Configuration.....	30
9.2.3.3 Other Encapsulation Configuration.....	30
9.2.4 Security Configuration.....	30
10. Security Considerations.....	31
10.1 IPsec.....	31
10.2 IS-IS Security.....	32
11. IANA Considerations.....	33
11.1 Port Assignments.....	33
11.2 Multicast Address Assignments.....	33
11.3 Encapsulation Method Support Indication.....	34
Normative References.....	35
Informative References.....	37
Acknowledgements.....	39
Authors' Addresses.....	40

1. Introduction

TRILL switches (also know as RBridges) are devices that implement the IETF TRILL protocol [RFC6325] [RFC7177] [RFC7780]. 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 [IS-IS] [RFC7176] link state routing with a TRILL header having a hop count.

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

This document specifies transmission of encapsulated TRILL data and TRILL IS-IS over IP (v4 or v6 [rfc2460bis]). so as to use an IP network as a TRILL link in a unified TRILL campus. Three encapsulations specified herein, two based on UDP and one based on TCP but provision is made to negotiate other encapsulations. TRILL over IP allows RBridges with IP connectivity to form a single TRILL campus, or multiple TRILL networks to be connected as a single TRILL campus via a TRILL over IP backbone.

The protocol specified in this document connects RBridge ports using transport over IP in such a way that the ports with IP connectivity appear to TRILL to be connected by a single multi-access link. If a set of more than two RBridge ports are connected via a single TRILL over IP link, each RBridge port in the set can communicate with every other RBridge port in the set.

To support the scenarios where RBridges are connected via IP paths (including those 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] for security.

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. Alternatively, where a common encapsulation is known to be supported by the TRILL switch ports on a link, they can simply be configured to always use that encapsulation.

This document updates [RFC7177] and [RFC7178] as described in Sections 5 and 11.3 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 link technology specific capabilities, in this case encapsulation methods supported 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 - See "encapsulation header".

encapsulation header - Protocol header or headers appearing between the IP Header and the TRILL Header. See Sections 4 and 5.

ESP - IPsec Encapsulating Security Protocol [RFC4303].

FGL - Fine Grained Label [RFC7172].

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

link - In TRILL, a link connects TRILL ports and may, for example, be a bridged LAN.

HKDF - Hash based Key Derivation Function [RFC5869].

MTU - Maximum Transmission Unit.

RBridge - Routing Bridge. An alternative term for a TRILL switch. [RFC6325] [RFC7780]

SNPA - Sub-Network Point of Attachment.

Sz - The campus wide MTU [RFC6325] [RFC7780].

TMCE - Traffic-Managed Controlled Environment, see Section 8.1.1.

TRILL - Transparent Interconnection of Lots of Links or Tunneled Routing in the Link Layer. The protocol specified in [RFC6325], [RFC7177], [RFC7780], and related RFCs.

TRILL switch - A device implementing the TRILL protocol.

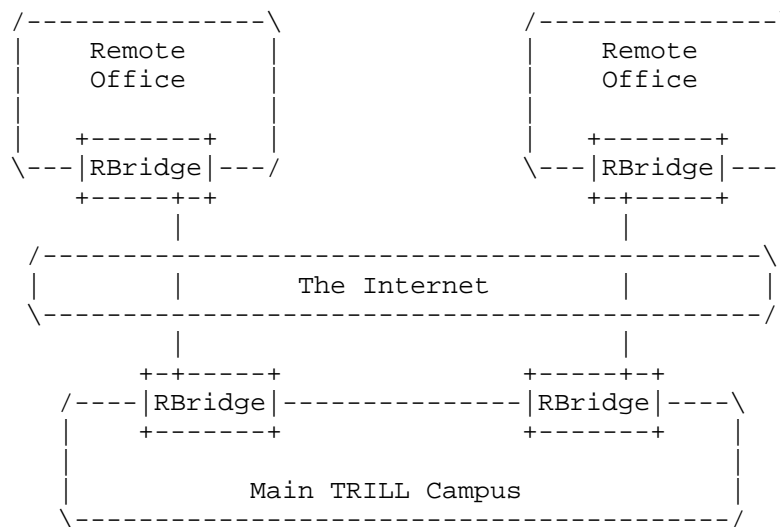
VNI - Virtual Network Identifier. In Virtual eXtensible Local Area Network (VXLAN) [RFC7348], the VXLAN Network Identifier.

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, as shown in the example below, 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

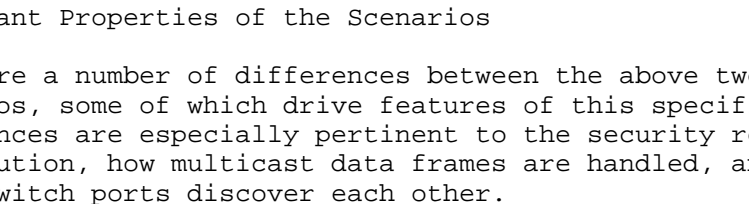
In the IP Backbone Scenario, as shown in the example below, TRILL over IP is used to connect a number of TRILL networks to form a single TRILL campus. For example, a TRILL over IP backbone could be used to connect multiple TRILL networks on different floors of a

Unlabeled TRILL Campus

TRILL Over IP Backbone

```

  graph TD
    subgraph "Unlabeled TRILL Campus"
      direction TB
      subgraph "TRILL Over IP Backbone"
          direction TB
          TopRow["+-----+-----+-----+"]
          BottomRow["+-----+-----+-----+"]
          TopRow --- V1["|"]
          TopRow --- V2["|"]
          TopRow --- V3["|"]
          V1 --- BottomRow
          V2 --- BottomRow
          V3 --- BottomRow
      end
    end
  
```



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.

Margaret Cullen, et al [Page 7]

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.

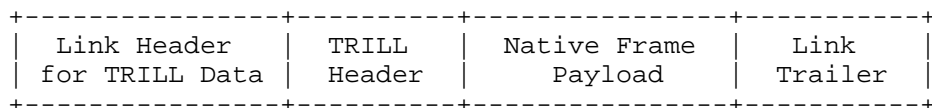
4. TRILL Packet Formats

To support TRILL 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

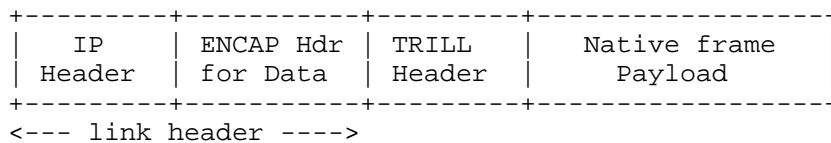
In TRILL over IP, we use an IP (v4 or v6) header followed by an encapsulation header, such as UDP, as the link header. (On the wire, the IP header will normally be preceded by the lower layer header of a protocol that is carrying IP; however, this does not concern us at the level of this document.)

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. Those encapsulations specified in this document are further detailed in Section 5. In the general specification below, those encapsulation fields will be represented as "ENCAP Hdr".

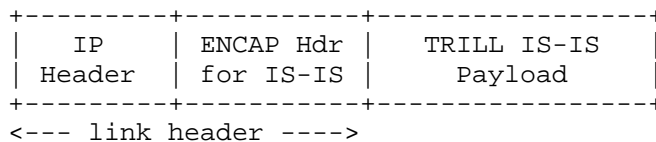
4.2.1 Without Security

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

TRILL Data Packet



TRILL IS-IS Packet

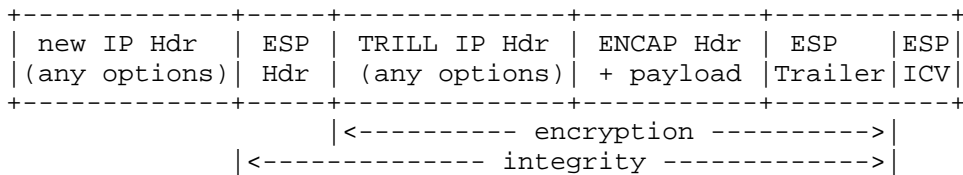


As discussed above and further specified in Section 5, the ENCAP Hdr indicates whether the packet is TRILL Data or IS-IS.

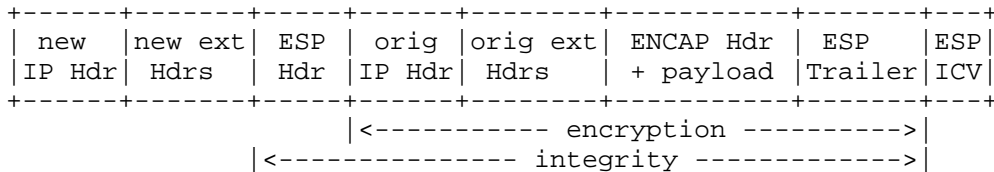
4.2.2 With Security

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 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:

With IPv4:



With IPv6:

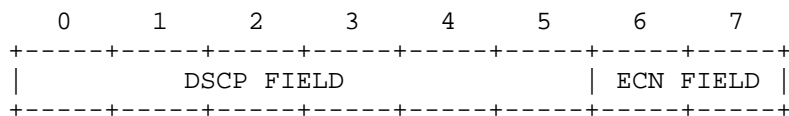


As shown above, IP Header options are considered part of the IPv4 Header but are extensions ("ext") of the IPv6 Header. For further information on the IPsec ESP Hdr, Trailer, and ICV, see [RFC4303] and Section 7 below. "ENCAP Hdr + payload" is the encapsulation header (Section 5) and TRILL data or the encapsulation header and 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 Differentiated Services Code Point (DSCP [RFC2474] [RFC3168]) in the IP 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. See [TRILLECN].)



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 [RFC7780]). (Typically TRILL IS-IS is configured to use 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, mapping the DSCP in this inner IP packet to the outer IP Header with the default for that mapping being to copy the DSCP without change.

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 priority 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 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 will be 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, including TRILL switches, establish adjacency through the exchange of Hello PDUs on a link [RFC7176] [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 network, 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:

```

    0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   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 addresses on that link, not MAC stations, even if the TRILL over IP Link is being carried over Ethernet. Therefore conflict on the link between a real MAC address and a TRILL over IP synthetic SNPA (MAC) address 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 exception of IS-IS MTU-probe and MTU-ack PDUs used to establish adjacency).

Alternatively, the network operator can pre-configure a TRILL over IP port to use a particular encapsulation chosen for their particular network's needs and port capabilities. That encapsulation is then used for all TRILL Data and IS-IS packets on ports so configured. This is expected to frequently be the case for a managed campus of TRILL switches.

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

An encapsulation must provide a method to distinguish TRILL Data packets and TRILL IS-IS packets or it is not useful for TRILL. In addition, the following criteria can be helpful in choosing between different encapsulations:

- a) Fast path support - For most 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 the 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 preferred.
- c) Robust fragmentation and re-assembly - The 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 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 a significant 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 support 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 unless the port is configured to always use some other encapsulation).

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 can be 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 a TRILL over IP port is using an encapsulation different from that in which Hellos are being exchanged, it is RECOMMENDED that BFD [RFC7175] or some other protocol that confirms adjacency using TRILL Data packets be used. As provided in [RFC7177] adjacency is not actually obtain until such confirmatory protocol succeeds.

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 by the receiver, that packet 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 IP 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 a broadcast link. In a well-configured network one would expect all of the IP ports involved to support the same encapsulation(s); but, for example, if P1 supports multiple encapsulations, it is possible that P2 and P3, do not have an encapsulation in common that is supported by P1. [IS-IS] can handle such non-transitive adjacencies that are reported as specified in [RFC7177]. This is generally done, albeit with reduced efficiency, by forwarding through the designated RBridge (router) on the link. Thus it is RECOMMENDED that all TRILL over IP ports on an IP link be configured to support one encapsulation in common that has good fast path support.

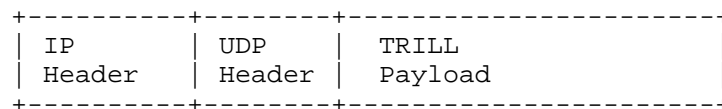
If serial IP unicast is being used by P1, it can use different encapsulations for different transmissions.

If native IP multicast is available for use by P1, it can send one transmission per encapsulation method by which it has a disjoint set of adjacencies on the link. If the transmitting port has adjacencies with overlapping sets of ports that are adjacent using different encapsulations, use of native multicast with different encapsulations may result in packet duplication. It would always be possible to use

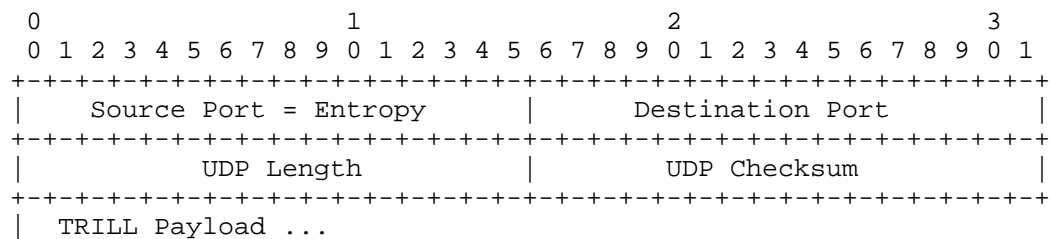
native IP multicast for one encapsulation for which the transmitting port has adjacencies, perhaps the encapsulation for which it has the largest number of adjacencies, and serially unicast to other receivers. These considerations are the reason that a TRILL over IP port MUST discard any packet received with an encapsulation for which it has not established an adjacency with the transmitter. Otherwise, packets would be further 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. This provides simple and direct access by TRILL to the native datagram service of IP.



Where the UDP Header is as follows:



Source Port - see Section 8.3

Destination Port - indicates TRILL Data or IS-IS, see Section 11.1.

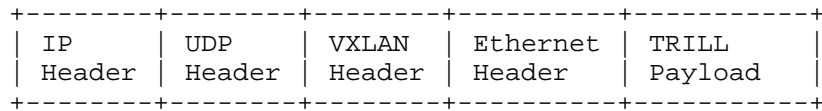
UDP Length - as specified in [RFC0768]

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.

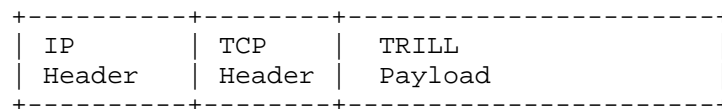


The outer UDP uses a destination port number indicating VXLAN and the outer UDP source port MAY be used for entropy as with native encapsulation (see Section 8.3). The VXLAN header after the outer UDP header adds a 24 bit Virtual Network Identifier (VNI). The Ethernet header after the VXLAN header and before the TRILL header consists of source MAC address, destination MAC address, and Ethertype. The Ethertype distinguishes TRILL Data from TRILL IS-IS. The destination and source MAC addresses in this Ethernet header are not used.

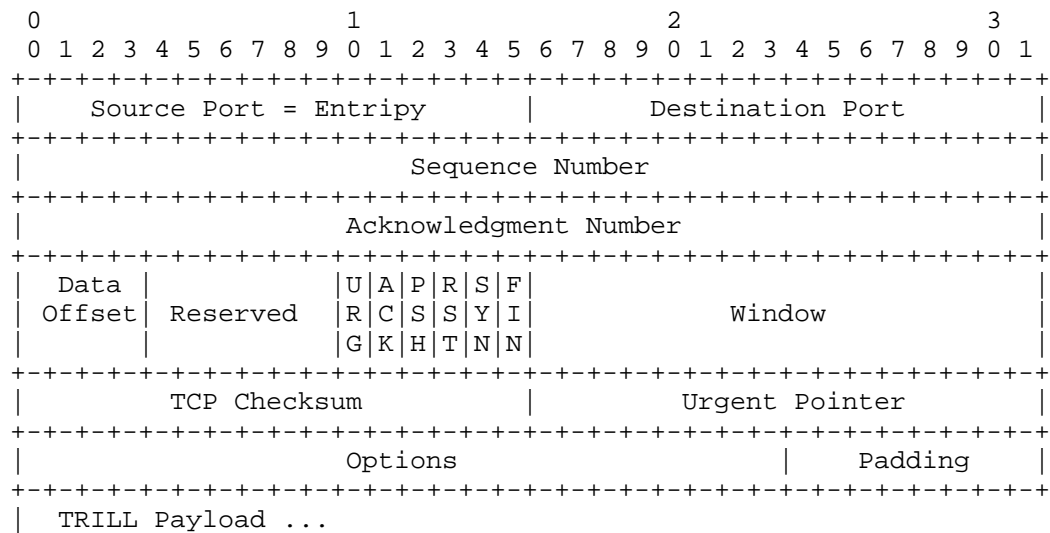
A TRILL over IP port using VXLAN encapsulation by default uses a VNI of 1 for TRILL IS-IS traffic and a VNI of 2 for TRILL data traffic but can be configured as described in Section 9.2.3.1 to use some other fixed VNIs or to map from VLAN/FGL to VNI.

5.6 TCP Encapsulation

TCP may be used for TRILL over IP as specified below. Use of TCP is convenient to provide congestion control (see Section 8.1) and reduced packet loss but is likely to cause substantial additional jitter and delay compared with a UDP based encapsulation.



Where the TCP Header is as follows:



Source Port - see Section 8.3

Destination Port - indicates TRILL Data or IS-IS, see Section 11.1.

Other TCP Header Fields - as specified in [RFC0793]

The TRILL Payload starts with the TRILL Header (not including the TRILL Ethertype) for TRILL Data packets and starts with the 0x83 Intradomain Routeing Protocol Discriminator byte (thus not including the L2-IS-IS Ethertype) for TRILL IS-IS packets.

5.7 Other Encapsulations

It is anticipated that additional TRILL over IP encapsulations will be specified in future documents and allocated a link technology specific flag bit as per Section 11.3. A primary consideration for whether it is worth the effort to specify use of an encapsulation by TRILL over IP is whether it has 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 (those with the M bit in the TRILL Header set to zero) 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 can be sent to them, but may be 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 two 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 TRILL Data packets transmission, since IS-IS security does not provide any protection to data packets, and SHOULD be fully established before any TRILL IS-IS packet transmission other than IS-IS Hello or MTU PDUs.

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 SHOULD 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. P1-System-ID and P2-System ID are the six byte System IDs of the two TRILL RBridges, and P1-Port and P2-Port are the TRILL Port IDs [RFC6325] of 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.

With [RFC5310] there could be multiple keys identified with 16-bit key IDs. The key ID when an IS-IS key is in use is transmitted in an IKEv2 ID_KEY_ID identity field [RFC7296] with Identification Data length of 2 bytes (Payload Length 6 bytes). The Key ID of the IS-IS-key is used to identify the IKEv2 shared secret derived as above that is actually used. ID_KEY_ID identity field(s) of other lengths MAY occur but their use is beyond the scope of this document.

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 an IS-IS shared key MUST expire within a lifetime no longer than the IS-IS-shared key from which they were derived. When the IKEv2 shared secret key expires, or earlier, the IKEv2 Security Association must be rekeyed using a new shared secret derived from a new IS-IS shared key.

IKEv2 with certificate based security MAY be used but details of certificate contents and use policy for this application of IKEv2 are beyond the scope of this document.

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 keying determined as in Section 7.1.1 is used for multi-destination TRILL traffic, which is unicast.

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 multi-destination TRILL traffic, which is unicast.

The case of a TRILL over IP port configured as broadcast and using native multicast is beyond the scope of this document. For security as provided in this document, multicast is handled via serial unicast.

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

This subsection discusses TRILL over UDP congestion considerations. These are applicable to the UDP based TRILL over IP encapsulation headers specified in detail in this document. Other encapsulations would likely have different congestion considerations and, in particular, the TCP encapsulation specified in Section 5.6 does not need congestion control beyond that provided by TCP. Congestion considerations for additional TRILL encapsulations will be provided in the document specifying the encapsulation.

One motivation for including UDP or TCP as the outermost part of a TRILL over IP encapsulation header is to improve the use of multipath such as Equal Cost Multi-Path (ECMP) in cases where traffic is to traverse routers that are able to hash on Port and IP address through addition of entropy in the source port (see Section 8.3). 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.

Section 3.1.11 of [RFC8085] discusses the congestion considerations for design and use of UDP tunnels; this is important because other flows could share the path with one or more UDP tunnels, necessitating congestion control [RFC2914] to avoid destructive interference.

The default initial determination of the TRILL over IP encapsulation to be used is through the exchange of TRILL IS-IS Hellos. This is a low bandwidth process. Hellos are not permitted to be sent any more often than once per second, and so are very unlikely to cause congestion. Thus no additional controls are needed for Hellos even if sent, as is the default, over UDP.

Congestion has potential impacts both on the rest of the network containing a UDP flow and on the traffic flows using the UDP encapsulation. These impacts depend upon what sort of traffic is carried in UDP, as well as the path it follows. The UDP based TRILL over IP encapsulations specified in this document do not provide any congestion control and are transmitted as regular UDP packets.

The two subsections below discuss congestion for TRILL over IP

traffic with UDP based encapsulation headers in traffic-managed controlled environments (TMCE, see [RFC8086]) and other environments.

8.1.1 Within a TMCE

Within a TMCE, that is, an IP network that is traffic-engineered and/or otherwise managed, for example via use of traffic rate limiters, to avoid congestion, UDP based TRILL over IP encapsulation headers are appropriate for carrying traffic that is not known to be congestion controlled. In such cases, operators of TMCE networks avoid congestion by careful provisioning of their networks, rate-limiting of user data traffic, and traffic engineering according to path capacity.

When TRILL over IP using a UDP based encapsulation header carries traffic that is not known to be congestion controlled in a TMCE network, the traffic path **MUST** be entirely within that network, and measures **SHOULD** be taken to prevent the traffic from "escaping" the network to the general Internet. Examples of such measures are:

- o physical or logical isolation of the links carrying the traffic from the general Internet and
- o deployment of packet filters that block the UDP ports assigned for TRILL over IP.

8.1.2 In Other Environments

Where UDP based encapsulation headers are used in TRILL over IP in environments other than those discussed in Section 8.1.1, specific congestion control mechanisms are commonly needed. However, if the traffic being carried by the TRILL over IP link is already congestion controlled and the size and volatility of the TRILL IS-IS link state database is limited, then specific congestion control may not be needed. See [RFC8085] Section 3.1.11 for further guidance.

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 subsequently sent to a next hop RBridge 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. (Or this might occur through a cycle of TRILL campuses.) The packet would 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 TRILL packets over any 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, for example with 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 UDP Source Port as an entropy field. (This idea is also introduced in [RFC8086].) 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; this range **SHOULD** be the ephemeral port range (49152-65535) except that, if there are middleboxes in the path (see Section 8.5), it **MUST** be possible to configure the range of different Source Port values to a sufficiently small range to avoid disrupting connectivity.

8.4 MTU Considerations

In TRILL each RBridge 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 the largest LSPs.

A method of determining `originatingLSPBufferSize` for an RBridge with one or more TRILL over IP ports is described in [RFC7780]. 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 RBridge 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 RBridge's `originatingLSPBufferSize`.

Because TRILL sets the minimum values of *Sz* at 1,470 bytes, R Bridges will not constrain LSPs or other TRILL IS-IS PDUs to a size smaller than that. Therefore there may be TRILL over IP links that require fragmentation to be enabled to accommodate such PDUs. When fragmentation is enabled, the effective link MTU from the TRILL point of view is larger than the RBridge port to RBridge port path MTU from the IP point of view. Path MTU discovery [RFC4821] should be useful in determining the IP MTU between a pair of RBridge ports with IP connectivity.

TRILL IS-IS MTU PDUs, as specified in Section 5 of [RFC6325] and in [RFC7177], can be used to obtain added assurance of the MTU of a link. An appropriate time to confirm MTU, or re-discover it if it has changed, is when an RBridge notices topology changes in a path that is in use for TRILL over IP due to LSP updates it receives; however, MTU can change at other times. For example, two RBridge ports are connected by a bridged LAN, topology or configuration changes within that bridged LAN could change the MTU between those RBridge ports.

For further discussion of these issues, see [IntareaTunnels].

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 for the usage of the zero UDP Checksum in a UDP tunnel protocol are detailed in [RFC6936]. These requirements apply to the UDP based TRILL over IP encapsulations specified herein (native and VXLAN), which are applications of UDP tunnel.

Besides the Checksum, the Source Port number of a UDP or TCP based ENCAP Hdr 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 or TCP 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 session. This provides the tunnel protocol application with the "NAT-pass-through" function. NAPT allows multiple internal hosts to share a single public IP address. The Source Port number, is used as the demultiplexer of the multiple internal hosts.

However, the above NAPT behavior conflicts with the behavior that the Source Port number is used as an entropy (See Section 8.3). Hence, the network operator MUST ensure the TRILL switch ports sending through local or remote NAPT middleboxes limit the entropy usage of the Source Port number, possibly to a single value.

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 TRILL over IP Differentiated Services Code Point (DSCP [RFC2474]) can be configured. If supported, mapping from an inner DSCP code point, when the TRILL payload is IP, to the outer TRILL over IP DSCP can be configured. (See Section 4.3.)

Each TRILL over IP port has a list of acceptable encapsulations it will use as the basis of adjacency. By default this list consists of one entry for native encapsulation (see Section 7). Additional encapsulations MAY be configured and native encapsulation MAY be removed from this list by configuration. 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).

Regardless of whether they will send IP multicast, TRILL over IP

ports emit appropriate IGMP (IPv4 [RFC3376]) or MLD (IPv6 [RFC2710]) packets unless configured not to do so. These are sent for the IP multicast group the port would use if it sent IP multicast.

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 multicast address for multicasting 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. IP packets received from a source IP address not on the list are discarded.

If this list of destination IP addresses is empty, the port is disabled.

9.2.3 Encapsulation Specific Configuration

Specific TRILL over IP encapsulation methods may provide for further configuration as specified below.

9.2.3.1 UDP and TCP Source Port

As discussed above, the UDP based encapsulation (Sections 5.4 and 5.5) and the TCP encapsulation (Section 5.6) start with a header containing a source port number that can be used for entropy (Section 8.3). The range of source port values used defaults to the ephemeral port range (49152-65535) but can be configured to any other range including to a single value.

9.2.3.2 VXLAN Configuration

A TRILL over IP port using VXLAN encapsulation can be configured with non-default VXLAN Network Identifiers (VNIs) that are used in that field of the VXLAN header for all TRILL IS-IS and TRILL Data packets sent using the encapsulation and required in those received using the encapsulation. The default VNI is 1 for TRILL IS-IS and 2 for TRILL Data. A TRILL packet received with the an unknown VNI is discarded.

A TRILL over IP port using VXLAN encapsulation can also be configured to map the Inner.VLAN of a TRILL Data packet being transported to the value it places in the VNI field and/or to copy the Inner.FGL [RFC7172] of a TRILL Data packet to the VNI field.

9.2.3.3 Other Encapsulation Configuration

Additional encapsulation methods, beyond those specified in this document, are expected to be specified in future documents and may require further configuration.

9.2.4 Security Configuration

A TRILL over IP port can be configured, for the case where IS-IS security [RFC5310] is in use, as to whether or not IPsec must be fully established and used for any TRILL IS-IS transmissions other than IS-IS Hello or MTU PDUs (see Section 7). There may also be configuration whose details are outside the scope of this document concerning certificate based IPsec or use of shared keys other than IS-IS based shared key or how to select what IS-IS based shared key to use.

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 3 above.

For communication between end stations in a TRILL campus, security may be 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 bridges or 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 would protect 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. It is anticipated that edge-to-edge TRILL security will be covered in future documents.

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 to protect any particularly sensitive information from unintended disclosure. Such end station to end station security is generally beyond the scope of TRILL

If VXLAN encapsulation is used, the unused Ethernet source and destination MAC addresses mentioned in Section 5.5, provide a 96 bit per packet side channel.

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 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 Ports in the Service Name and Transport Protocol Port Number Registry [PortRegistry] for TRILL IS-IS and TRILL Data as shown below.

```
Service Name: TRILL-IS-IS
Transport Protocol: udp, tcp
Assignee: iesg@ietf.org
Contact: chair@ietf.org
Description: Transport of TRILL IS-IS control PDUs.
Reference: [this document]
Port Number: (TBD1)
```

```
Service Name: TRILL-data
Transport Protocol: udp, tcp
Assignee: iesg@ietf.org
Contact: chair@ietf.org
Description: Transport of TRILL Data packets.
Reference: [this document]
Port Number: (TBD2)
```

11.2 Multicast Address Assignments

IANA is requested to assign one IPv4 and one IPv6 multicast address, as shown below, which correspond to both the All-RBridges and All-IS-IS-RBridges multicast MAC addresses that have been 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[233.252.1.32]	TBD4[FF0X:0:0:0:0:0:0:BAC1]

The hex digit "X" in the IPv6 variable scope 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 indented 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: Expert Review
 Reference: [this document]

Flag	Meaning	Reference
-----	-----	-----
0xFD0	Native encapsulation supported	[this document]
0xFD1	VXLAN encapsulation supported	[this document]
0xFD2	TCP encapsulation supported	[this document]
0xFD3-0xFF7	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>>.
- [RFC0793] - Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<http://www.rfc-editor.org/info/rfc793>>.
- [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>>. <<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>>.
- [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, DOI 10.17487/RFC7175, May 2014, <<http://www.rfc-editor.org/info/rfc7175>>.
- [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>>.

[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

[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>>.

[RFC4821] - Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <<http://www.rfc-editor.org/info/rfc4821>>.

[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>>.

[RFC8085] - Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <<http://www.rfc-editor.org/info/rfc8085>>.

[RFC8086] - Yong, L., Ed., Crabbe, E., Xu, X., and T. Herbert, "GRE-in-UDP Encapsulation", RFC 8086, DOI 10.17487/RFC8086, March 2017, <<http://www.rfc-editor.org/info/rfc8086>>.

[IntareaTunnels] - J. Touch, M. Townsley, "IP Tunnels in the Internet Architecture", draft-ietf-intarea-tunnels, work in progress.

[LinkSec] - Eastlake, D., D. Zhang, "TRILL: Link Security", draft-eastlake-trill-link-security, work in progress.

[TRILLECN] - Eastlake, D., B. Briscoe, "TRILL: ECN (Explicit Congestion Notification) Support", draft-ietf-trill-ecn-support, work in progress.

[PortRegistry] - <https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml>

Acknowledgements

The following people have provided useful feedback on the contents of this document: Sam Hartman, Adrian Farrel, Radia Perlman, Ines Robles, Joe Touch, Mohammed Umair, Lucy Yong.

Some of the material in this document is derived from [RFC8085] and [RFC8086].

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

Authors' Addresses

Margaret Cullen
Painless Security
14 Summer Street, Suite 202
Malden, MA 02148
USA

Phone: +1-781-605-3459
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
Huawei Technologies

Email: dacheng.zhang@huawei.com

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Margaret Cullen
Painless Security
Donald Eastlake
Mingui Zhang
Dacheng Zhang
Huawei
May 21, 2018

Expires: November 20, 2018

TRILL (Transparent Interconnection of Lots of Links)
Over IP Transport
<draft-ietf-trill-over-ip-17.txt>

Abstract

The TRILL (Transparent Interconnection of Lots of Links) 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 specifies transmission of encapsulated TRILL data and IS-IS over IP (v4 or v6) transport. so as to use an IP network as a TRILL link in a unified TRILL campus. This document 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 authors or the TRILL Working Group 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.....	6
3. Use Cases for TRILL over IP Transport.....	8
3.1 Remote Office Scenario.....	8
3.2 IP Backbone Scenario.....	8
3.3 Important Properties of the Scenarios.....	9
3.3.1 Security Requirements.....	9
3.3.2 Multicast Handling.....	10
3.3.3 Neighbor Discovery.....	10
4. TRILL Packet Formats.....	11
4.1 General Packet Formats.....	11
4.2 General TRILL Over IP Packet Formats.....	12
4.2.1 Without Security.....	12
4.2.2 With Security.....	12
4.3 QoS Considerations.....	13
4.4 Broadcast Links and Multicast Packets.....	15
4.5 TRILL Over IP Transport IS-IS SubNetwork Point of Attachment.....	15
5. TRILL over IP Transport Encapsulation Formats.....	17
5.1 Encapsulation Considerations.....	17
5.2 Encapsulation Agreement.....	18
5.3 Broadcast Link Encapsulation Considerations.....	19
5.4 Native Encapsulation.....	20
5.4.1 IPv4 UDP Checksum Considerations.....	21
5.4.2 IPv6 UDP Checksum Considerations.....	22
5.5 VXLAN Encapsulation.....	24
5.6 TCP Encapsulation.....	25
5.6.1 TCP Connection Establishment.....	26
5.7 Other Encapsulations.....	27
6. Handling Multicast.....	28
7. Use of IPsec and IKEv2.....	29
7.1 Keying.....	29
7.1.1 Pairwise Keying.....	29
7.1.2 Group Keying.....	30
7.2 Mandatory-to-Implement Algorithms.....	31
8. Transport Considerations.....	32
8.1 UDP Congestion Considerations.....	32
8.1.1 Within a TMCE.....	33
8.1.2 In Other Environments.....	33
8.2 Recursive Ingress.....	34
8.3 Fat Flows.....	34
8.4 MTU Considerations.....	35

Table of Contents (continued)

9. TRILL over IP Transport Port Configuration.....	37
9.1 Per IP Port Configuration.....	37
9.2 Additional per IP Address Configuration.....	37
9.2.1 Native Multicast Configuration.....	38
9.2.2 Serial Unicast Configuration.....	38
9.2.3 Encapsulation Specific Configuration.....	38
9.2.3.1 UDP Source Port.....	38
9.2.3.2 VXLAN Configuration.....	39
9.2.3.3 TCP Configuration.....	39
9.2.3.4 Other Encapsulation Configuration.....	39
9.2.4 Security Configuration.....	39
10. Security Considerations.....	40
10.1 IPsec.....	40
10.2 IS-IS Security.....	41
11. IANA Considerations.....	42
11.1 Port Assignments.....	42
11.2 Multicast Address Assignments.....	42
11.3 Encapsulation Method Support Indication.....	43
Normative References.....	45
Informative References.....	47
Appendix A: IP Security Choice.....	50
Acknowledgements.....	51
Authors' Addresses.....	52

1. Introduction

TRILL switches (also know as RBridges) are devices that implement the IETF TRILL protocol [RFC6325] [RFC7177] [RFC7780]. 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 [IS-IS] [RFC7176] link state routing and transmits data using a TRILL header that has a hop count.

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

This document specifies transmission of encapsulated IS-IS and TRILL data over IP (v4 or v6 [RFC8200]) transport. so as to use an IP network as a TRILL link in a unified TRILL campus. One mandatory to implement UDP based encapsulation is specified along with two optional to implement encapsulations, one using VXLAN (which is based on UDP), and one using TCP. Provision is made to negotiate other encapsulations. TRILL over IP transport allows RBridges with IP connectivity to form a single TRILL campus, or multiple TRILL networks to be connected as a single TRILL campus via a TRILL over IP transport backbone.

The protocol specified in this document connects RBridge ports using transport over IP transport in such a way that the ports with mutual IP connectivity appear to TRILL to be connected by a single multi-access link. If a set of more than two RBridge ports are connected via a single TRILL over IP transport link, each RBridge port in the set can communicate with every other RBridge port in the set.

To support the scenarios where RBridges are connected via IP paths (including those 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] as the mandatory to implement protocol for security (see appendix A).

To dynamically select a mutually supported TRILL over IP transport 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. Alternatively, where a common encapsulation is known to be fully supported by the TRILL switch ports on a link, those ports can simply be configured to always use that encapsulation.

This document updates [RFC7177] and [RFC7178] as described in Sections 5 and 11.3 by making adjacency between TRILL over IP transport ports dependent on having a fully supported method of

encapsulation in common and by redefining an interval of RBridge Channel protocol numbers to indicate link technology specific capabilities, in this case encapsulation methods supported for TRILL over IP transport.

2. Terminology

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

The following terms and acronyms have the meaning indicated:

DEI - Drop Eligibility Indicator. Part of QoS, see Section 4.3.

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

ENCAP Hdr - See "encapsulation header".

encapsulation header - Protocol header or headers appearing between the IP Header and the TRILL Header. See Sections 4 and 5.

ESP - IPsec Encapsulating Security Protocol [RFC4303].

FGL - Fine Grained Label [RFC7172].

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

link - In TRILL, a link connects TRILL ports and is transparent to TRILL data and IS-IS messages. It may, for example, be a bridged LAN.

HKDF - Hash based Key Derivation Function [RFC5869].

MTU - Maximum Transmission Unit.

PDU - Protocol Data Unit.

QoS - Quality of Service.

RBridge - Routing Bridge. An alternative term for a TRILL switch. [RFC6325] [RFC7780]

SNPA - Sub-Network Point of Attachment [IS-IS].

Sz - The campus wide MTU [RFC6325] [RFC7780].

TMCE - Traffic-Managed Controlled Environment, see Section 8.1.1.

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

[RFC7177], [RFC7780], and related RFCs.

TRILL switch - A device implementing the TRILL protocol.

VNI - Virtual Network Identifier. In Virtual eXtensible Local Area Network (VXLAN) [RFC7348], the VXLAN Network Identifier.

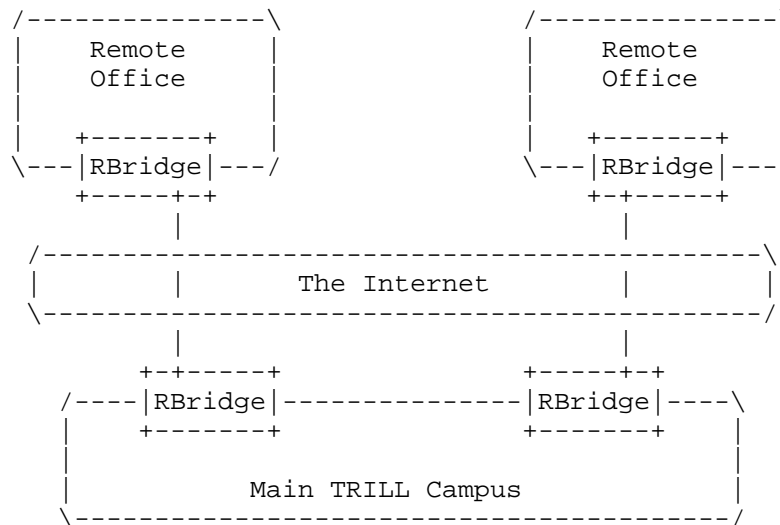
3. Use Cases for TRILL over IP Transport

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, as shown in the example below, 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 the 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.

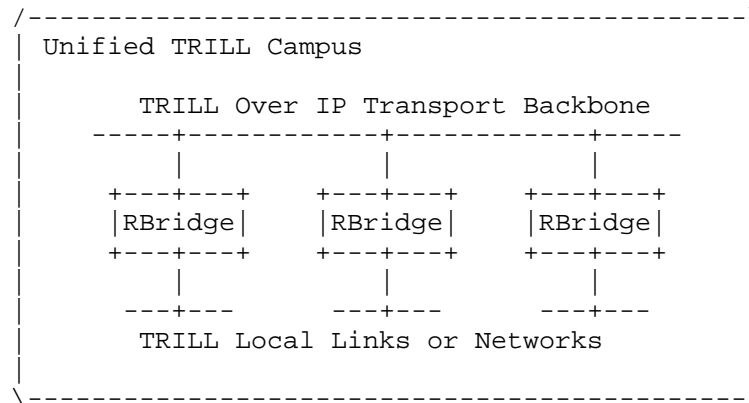
In this use case, the TRILL over IP transport link will often cross logical and physical IP networks that do not support TRILL, are not under the same administrative control as the TRILL campus, and whose level of physical security is unknown.



3.2 IP Backbone Scenario

In the IP Backbone Scenario, as shown in the example below, TRILL over IP transport is used to connect a number of TRILL networks to

form a single TRILL campus. For example, a TRILL over IP transport 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 transport link, and the IP link(s) used by TRILL over IP transport are typically under the same administrative control as the rest of the TRILL campus and might or might not be physically secure.



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 transport 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. The integrity of TRILL routing and the TRILL campus depend on protection of RBridges from compromise; if similar security can be extended to the links between RBridges, 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 transport may run over a

network that is not under the same administrative control as the TRILL network. It may appear to nodes on the network 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 can be addressed by using IPsec for both 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 transport link. If so, it can be used to send IS-IS PDUs 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 more 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 IS-IS control messages and multicast data packets that are transmitted between the Remote Office and the TRILL campus will need to be serially transmitted by IP unicast, as discussed later in this document.

3.3.3 Neighbor Discovery

In the IP Backbone Scenario, where IP multicast is supported, TRILL switches that use TRILL over IP transport can use the normal 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 usually need to be established before 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.

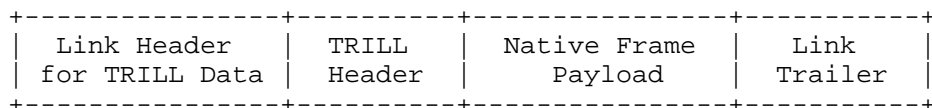
4. TRILL Packet Formats

To support TRILL two types of packets are transmitted between TRILL switches: IS-IS packets and TRILL Data packets.

Section 4.1 describes general packet formats for TRILL data and IS-IS independent of link technology. Section 4.2 specifies general TRILL over IP transport 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 IS-IS Hello SubNetwork Point of Attachment (SNPA) considerations for IP transport.

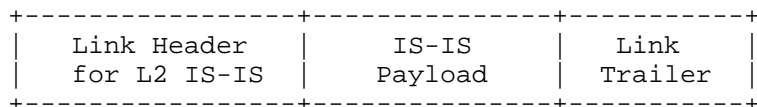
4.1 General Packet Formats

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).

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 IS-IS from TRILL Data. For example, over Ethernet, the Link Header for TRILL Data ends with the TRILL Ethertype while the Link Header for IS-IS ends with the L2-IS-IS Ethertype; on the other hand, over PPP, there are no Ethernets in the Link Header but different PPP protocol code points are included that distinguish IS-IS from TRILL Data.

4.2 General TRILL Over IP Packet Formats

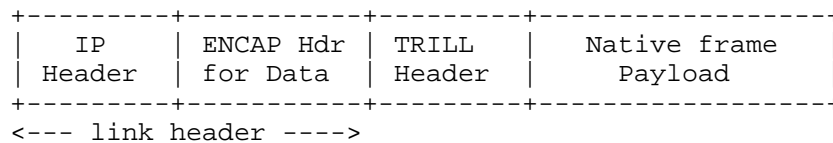
In TRILL over IP transport, we use an IP (v4 or v6) header followed by an encapsulation header, such as UDP, as the link header. (On the wire, the IP header will normally be preceded by the lower layer header of a protocol that is carrying IP; however, this does not concern us at the level of this document.)

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

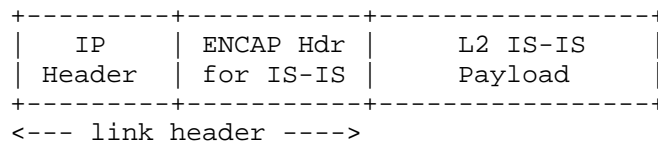
4.2.1 Without Security

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

TRILL Data Packet



L2 IS-IS Packet



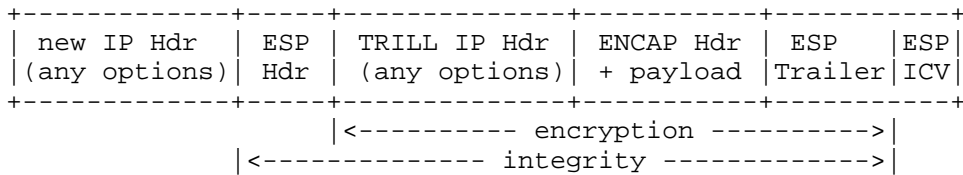
As discussed above and further specified in Section 5, the ENCAP Hdr indicates whether the packet is TRILL Data or IS-IS.

4.2.2 With Security

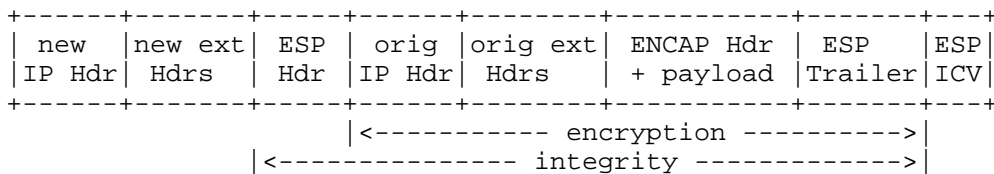
The mandatory to implement TRILL over IP transport link security is IPsec Encapsulating Security Protocol (ESP) in tunnel mode [RFC4303] (see Appendix A). Since TRILL over IP transport always starts with an IP Header (on the wire this appears after any lower layer header that might be required), the modifications for IPsec are independent of the TRILL over IP transport ENCAP Hdr that occurs after that IP Header. ENCAP headers specified in this document are UDP, VXLAN, and

TCP. The resulting packet formats are as follows for IPv4 and IPv6:

With IPv4:



With IPv6:

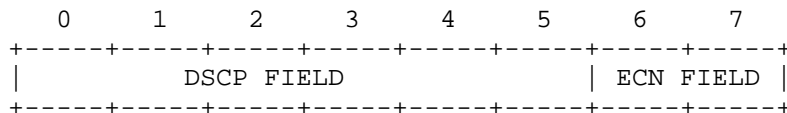


As shown above, IP Header options are considered part of the IPv4 Header but are extensions ("ext") of the IPv6 Header. For further information on the IPsec ESP Hdr, Trailer, and ICV, see [RFC4303] and Section 7 below. "ENCAP Hdr + payload" is the encapsulation header (Section 5) and TRILL data or the encapsulation header and 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 transport RBridge port (see, for example, Section 1.1.3 of [RFC7296]).

4.3 QoS Considerations

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



DSCP: Differentiated Services Codepoint

ECN: Explicit Congestion Notification

Although recommendations are provided below for mapping from TRILL

priority to DSCP, behavior for various DSCP values on the general Internet is not predictable. The default mapping below is appropriate where the TRILL campus is under the control of a network manager or consists of islands connected by an Internet Service Provider where that manager and/or provider support the DSCPs below to provide the QoS indicated.

Within a TRILL switch, QoS is determined (1) by configuration for IS-IS packets and (2) by a three bit (0 through 7) priority field and a Drop Eligibility Indicator (DEI) bit (see Sections 8.2 and 7 of [RFC7780]) for TRILL Data packets. (Typically IS-IS is configured to use one of the highest two priorities depending on the particular IS-IS PDU.) The QoS 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 Ethernet and thus might be a bridged LAN, QoS is commonly encoded into an Outer.VLAN tag's priority and DEI fields.

TRILL over IP transport 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, mapping the DSCP in this inner IP packet to the DSCP in the outer IP Header with the default for that mapping being to copy the DSCP without change.

The default TRILL priority and DEI to DSCP mapping, which may be configured per TRILL over IP transport 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 priority 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, as it is in [802.1Q].

TRILL Priority	DEI	DSCP Field (Binary/decimal)
-----	---	-----
0	0/1	000000 / 0
1	0/1	-TBD0- / TBD0
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

RFC Editor: Please change the TBD0 DSCP for TRILL Priority 1 in the above table and below text to the DSCP value that is recommended for the Lower Effort PHB (LE PHB) by draft-ietf-tsvwg-le-phb [LEphb] draft when that draft is published as an RFC and delete this note.

The above all follow the recommended DSCP values from [RFC2474] except for the placing of priority 1 below priority 0, as specified in [802.1Q], and for the DSCP value of TBD0 binary for low effort as recommended in [LEphb].

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 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 will be 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 IS-IS respectively. For details on TRILL over IP transport handling of multicast, see Section 6.

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

IS-IS routers, including TRILL switches, establish adjacency through the exchange of Hello PDUs on a link [RFC7176] [RFC7177]. The Hellos transmitted out of 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 network, the SNPA is the port MAC address.)

In IS-IS Hello PDUs on a TRILL over IP transport 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 ease 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:

```
0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```



```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   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 as specified in this document, TRILL sees only IP ports on that link, not MAC stations, even if the TRILL over IP transport link is being carried over Ethernet. Therefore conflicts on the link between a real MAC address and a TRILL over IP transport synthetic SNPA (MAC) address are impossible.

5. TRILL over IP Transport Encapsulation Formats

There are a variety of TRILL over IP transport encapsulation formats possible. There are two levels of support for an encapsulation by a TRILL over IP transport port as follows:

limited support - Limited support for an encapsulation enables the exchange of TRILL IS-IS Hellos and other adjacency related PDUs, including the information needed to determine what, if any, fully supported encapsulation the port has in common with other ports.

full support - Full support by a TRILL over IP transport port for an encapsulation means the port enables use of that encapsulation for data and all control messages if the encapsulation is negotiated with another such port on the link.

By default TRILL over IP transport adopts a hybrid encapsulation approach. All TRILL over IP transport ports **MUST** implement limited support for native encapsulation (see Section 5.4). Although native encapsulation does not typically have good fast path support, as a lowest common denominator it can be used for 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 fully supported encapsulation used for all TRILL Data packets and all other TRILL IS-IS PDUs (with the exception of IS-IS MTU-probe and MTU-ack PDUs used to establish adjacency which also use native encapsulation by default).

Alternatively, the network operator can pre-configure a TRILL over IP transport port to always use a particular encapsulation chosen for their particular network's needs and port capabilities. That encapsulation is then used for all TRILL Data and IS-IS packets, including Hellos, on ports so configured. This is expected to frequently be the case for a managed campus of TRILL switches.

Section 5.1 discusses general considerations for the TRILL over IP transport encapsulation format. Section 5.2 discusses encapsulation agreement. Section 5.3 discusses broadcast link encapsulation considerations. Section 5.4 and subsequent subsections discuss particular encapsulations.

5.1 Encapsulation Considerations

An encapsulation must provide a method to distinguish TRILL Data packets and TRILL IS-IS packets or it is not useful for TRILL. In

addition, the following criteria can be helpful in choosing between different encapsulations for full support:

- a) Fast path support - For most applications, it is highly desirable to be able to encapsulate/decapsulate TRILL over IP transport at line speed. Thus a format where existing or anticipated fast path hardware can do that is best. This is commonly the dominant consideration.
- b) Ease of multi-pathing - The IP path between TRILL over IP transport ports may include equal cost multipath routes internal to the IP link so a method of encapsulation that provides variable fields available for fast path hardware multi-pathing is preferred.
- c) Robust fragmentation and re-assembly - Fragmentation should generally be avoided; however, the 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 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 transport link, a checksum provided by the encapsulation may be a significant factor. Use of IPsec can also provide a strong integrity check.

5.2 Encapsulation Agreement

TRILL Hellos sent out of a TRILL over IP transport port indicate the encapsulations for which that port is offering full 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 transport, indicate support for different encapsulations, allowing support for up to 40 encapsulations to be specified. Full support for an encapsulation is indicated in the Hello PDU using the same mechanism by which support for an RBridge Channel protocol is indicated (see also section 11.3). Such full 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 unless the port is configured to always use some other encapsulation).

If, in a TRILL Hello on a TRILL over IP transport link, full support

is not indicated for any encapsulation, then the port from which it was sent is assumed to fully support native encapsulation only (see Section 5.4).

An adjacency can be formed between two TRILL over IP transport ports if the intersection of the sets of encapsulation methods they fully support is not null. If that intersection is null, then no adjacency is formed. In particular, for a TRILL over IP transport link, the adjacency state machine **MUST NOT** advance to the Report state unless the ports share a fully supported encapsulation [RFC7177]. If no such encapsulation is shared, the adjacency state machine remains in the state from which it would otherwise have transitioned to the Report state when an event occurs that would have transitioned it to the Report state.

If a TRILL over IP transport port is using an encapsulation different from that in which Hellos are being exchanged, it is **RECOMMENDED** that BFD [RFC7175] or some other protocol that confirms adjacency using TRILL Data packets be used. As provided in [RFC7177], adjacency is not actually obtained when such a confirmatory protocol is in use until that protocol succeeds.

If any TRILL over IP transport packet, other than an IS-IS Hello or MTU PDU in native encapsulation, is received in an encapsulation for which full support is not being indicated by the receiver, that packet **MUST** be discarded (see Section 5.3).

If there are two or more fully supported encapsulations in common between two adjacent ports for unicast or across all of the set of adjacent ports for multicast, a transmitter is free to choose whichever of those 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, that is to say, encapsulation usage could be asymmetric.

It is expected to be the normal case in a well-configured network that all the TRILL over IP transport ports connected to an IP link (i.e., an IP network) that are intended to communicate with each other fully support the same encapsulation(s).

5.3 Broadcast Link Encapsulation Considerations

To properly handle TRILL protocol packets on a TRILL over IP transport 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 IP multicast addressed packets.)

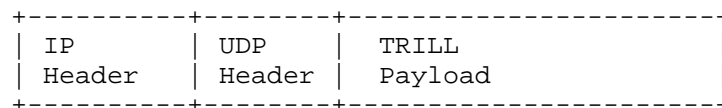
It is possible for the Hellos from a TRILL over IP transport port P1 to establish adjacency with multiple other TRILL over IP transport ports (P2, P3, ...) on a broadcast link. In a well-configured network one would expect all of the IP ports involved to fully support the same encapsulation; but, for example, if P1 fully supports multiple encapsulations, it is possible that P2 and P3, do not have an encapsulation in common that is also supported by P1. [IS-IS] can handle such non-transitive adjacencies that are reported as specified in [RFC7177]. This is generally done, albeit with reduced efficiency, by forwarding through the designated RBridge (router) on the link. Thus it is RECOMENDED that all TRILL over IP transport ports on an IP link be configured to fully support one encapsulation in common that has good fast path support.

If serial IP unicast is being used by P1, it MAY use different encapsulations for different transmissions.

If multiple IP multicast encapsulations are available for use by P1, it can send one transmission per encapsulation method by which it has a disjoint set of adjacencies on the link. If the transmitting port has adjacencies with overlapping sets of ports that are adjacent using different encapsulations, use of native multicast with different encapsulations may result in packet duplication. It would always be possible to use native IP multicast for one encapsulation or multiple encapsulations supported by non-overlapping sets of receiving ports for which the transmitting port has adjacencies, perhaps the encapsulation(s) for which it has the largest number of adjacencies, and serially unicast to other receivers. These considerations are the reason that a TRILL over IP transport port MUST discard any packet received with an encapsulation for which it has not established an adjacency with the transmitter. Otherwise, packets might be further duplicated.

5.4 Native Encapsulation

The mandatory to implement "native encapsulation" format of a TRILL over IP transport packet, when used without security, is TRILL over UDP as shown below. This provides simple and direct access by TRILL to the native datagram service of IP.



Where the UDP Header is as follows:

```

      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 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|   Source Port = Entropy   |   Destination Port   |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          UDP Length       |          UDP Checksum   |
+-----+-----+-----+-----+-----+-----+-----+-----+
|   TRILL Payload ...      |

```

Source Port - see Section 8.3.

Destination Port - indicates TRILL Data or IS-IS, see Section 11.1.

UDP Length - as specified in [RFC0768].

UDP Checksum - as specified in [RFC0768]. See discussion below.

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.

Note that if the mandatory to implement TRILL over IP transport security is in use, then traffic is not actually over UDP but rather over IPsec ESP. The authentication / integrity services provided protect against the processing of traffic by the wrong receiver even when the destination IP address / port is corrupted or the like and the confidentiality services provided by IPsec protect against compromise even if a receiver attempts to process packets not originally addressed to it.

5.4.1 IPv4 UDP Checksum Considerations

For UDP in IPv4, when a non-zero UDP checksum is used, the UDP checksum MUST be processed as specified in [RFC0768] and [RFC1122] for both transmit and receive. The IPv4 header includes a checksum that protects against misdelivery of the packet due to corruption of IP addresses. The UDP checksum potentially provides protection against corruption of the UDP header and TRILL payload. Disabling the use of checksums is a deployment consideration that should take into account the risk and effects of packet corruption.

When a port receives a TRILL over IP transport packet, the UDP checksum field MUST be processed. If the UDP checksum is non-zero, the port MUST verify the checksum before accepting the packet. By default, a TRILL over IP transport port SHOULD accept UDP packets with a zero checksum. A node MAY be configured to disallow zero

checksums per [RFC1122]; this may be done selectively, for instance, disallowing zero checksums from certain adjacent ports that are known to be sending over paths subject to packet corruption. If verification of a non-zero checksum fails, a port lacks the capability to verify a non-zero checksum, or a packet with a zero checksum was received and the port is configured to disallow, the packet MUST be dropped and an event MAY be logged.

5.4.2 IPv6 UDP Checksum Considerations

For UDP in IPv6, the UDP checksum MUST be processed as specified in [RFC0768] and [RFC8200] for both transmit and receive.

When UDP is used over IPv6, the UDP checksum is relied upon to protect both the IPv6 and UDP headers from corruption. As such, a default TRILL over IP transport port MUST perform UDP checksum; a traffic-managed controlled environment (TMCE) TRILL over IP transport port MAY be configured with UDP zero-checksum mode if the TMCE or a set of closely cooperating TMCEs (such as by network operators who have agreed to work together in order to jointly provide specific services) meet at least one of the following conditions:

- a. It is known (perhaps through knowledge of equipment types and lower-layer checks) that packet corruption is exceptionally unlikely and where the operator is willing to take the risk of undetected packet corruption.
- b. It is judged through observational measurements (perhaps of historic or current traffic flows that use a non-zero checksum) that the level of packet corruption is tolerably low and where the operator is willing to take the risk of undetected packet corruption.
- c. Carrying applications that are tolerant of misdelivered or corrupted packets (perhaps through higher-layer checksum, validation, and retransmission or transmission redundancy) where the operator is willing to rely on the applications using the tunnel to survive any corrupt packets.

The following requirements apply to a TMCE TRILL over IP transport port that uses UDP zero-checksum mode:

- a. Use of the UDP checksum MUST be the default configuration of all IPv6 TRILL over IP transport ports.
- b. The port implementation MUST comply with all requirements specified in Section 4 of [RFC6936] and with requirement 1 specified in Section 5 of [RFC6936].

- c. A receiving TRILL over IP transport port SHOULD only allow the use of UDP zero checksum mode for IPv6 that is sent to one of the two TRILL over IP UDP Destination Port numbers (see Section 11.1). The motivation for this requirement is possible corruption of the UDP Destination Port, which may cause packet delivery to the wrong UDP port. If that other UDP port requires the UDP checksum, the misdelivered packet will be discarded.
- d. It is RECOMMENDED that the UDP zero-checksum mode for IPv6 only be enabled for TRILL over IP transport ports with a configured set of possible adjacencies. Because TRILL data is discarded unless it is received from a source address with which an adjacency exists, the receiving TRILL over IP transport port will check the source IPv6 address and MUST check that the destination IPv6 address is appropriate if UDP zero-checksum is being used and discard any packet for which these checks fails.
- e. This document assumes there are no middleboxes in the path and thus does not cover restrictions on such middleboxes. Middlebox support is beyond the scope of this document.
- f. Measures SHOULD be taken to prevent IPv6 traffic with zero UDP checksums from "escaping" to the general Internet.
- g. IPv6 traffic with zero UDP checksums MUST be actively monitored for errors by the network operator. For example, the operator may monitor Ethernet-layer packet error rates.
- h. If a packet with a non-zero checksum is received, the checksum MUST be verified before accepting the packet regardless of port configuration to use UDP zero-checksum mode.

The above requirements do not change either the requirements specified in [RFC8200] or the requirements specified in [RFC6936].

The requirements to check the source and destination IPv6 addresses provide some mitigation for the absence of UDP checksum coverage of the IPv6 header. A TMCE that satisfies at least one of three conditions listed at the beginning of this section provides additional assurance.

TRILL over IP/UDP is suitable for transmission over lower layers in TMCEs that are allowed by the exceptions stated above. The rate of corruption of the inner IP packet on such networks is not expected to increase by comparison to TRILL traffic that is not encapsulated in UDP. Typically lower layers do provide some integrity checking such as the FCS (Frame Check Sequence) at the end of Ethernet packets. This design is in accordance with requirements 2, 3, and 5 specified in Section 5 of [RFC6936].

TRILL over IP/UDP does not accumulate incorrect transport-layer state as a consequence of IP/UDP header corruption. Such corruption may result in either packet discard or packet forwarding but the IP/UDP header is stripped at the end of each TRILL over IP transport hop between RBridges so errors cannot accumulate. Active monitoring of TRILL over IP/UDP traffic for errors is REQUIRED, as the occurrence of errors will result in some accumulation of error information outside the protocol for operational and management purposes. This design is in accordance with requirement 4 specified in Section 5 of [RFC6936].

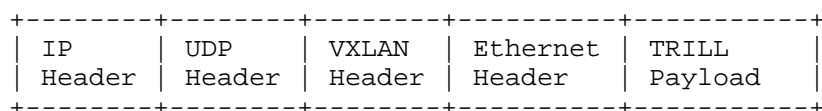
The remaining requirements specified in Section 5 of [RFC6936] are not applicable to TRILL over IP/UDP. Requirements 6 and 7 do not apply because TRILL over IP/UDP does not include a control feedback mechanism. Requirements 8-10 are middlebox requirements that do not apply to TRILL over IP/UDP ports and, in any case, middleboxes are out of scope for this document.

It is worth mentioning that the use of a zero UDP checksum should present the equivalent risk of undetected packet corruption when sending a similar packet using underlying Layer 2 link protocols in the cases where those protocols do not have a checksum.

In summary, a TMCE TRILL over IP/UDP is allowed to use UDP zero-checksum mode for IPv6 when the conditions and requirements stated above are met. Otherwise, the UDP checksum needs to be used for IPv6 as specified in [RFC0768] and [RFC8200].

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 8.3). UDP checksum considerations are the same as in 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. The destination and source MAC addresses

in this Ethernet header are not used.

A TRILL over IP port using VXLAN encapsulation by default uses a VNI of 1 for TRILL IS-IS traffic and a VNI of 2 for TRILL data traffic but can be configured as described in Section 9.2.3.1 to use some other fixed VNIs or to map from VLAN/FGL to VNI for data traffic.

5.6 TCP Encapsulation

TCP [RFC0793] may be used for TRILL over IP transport as specified below. Use of TCP is convenient to provide congestion control (see Section 8.1) and reduced packet loss but is likely to cause substantial additional jitter and delay compared with a UDP based encapsulation.

TCP supports only unicast communication. Thus, when TCP encapsulation is being used, multi-destination packets must be sent by serial unicast. Neighbor discovery cannot be done with TCP, so if discovery is to be supported at a TRILL over IP transport port (i.e., the set of potential adjacencies is not configured), Hellos must be sent with UDP native encapsulation. If a TRILL over IP transport port is configured to use TCP encapsulation for all traffic, a list of IP addresses that port might communicate with must be configured for the port (see Section 9).

All packets in a particular TCP stream MUST use the same DSCP value as discussed in [RFC7657]. Therefore a TCP connection is needed per QoS to be provided between TRILL switches. Contiguous sets of priority levels MAY be mapped into a single TCP connection with a single DSCP value. Lower priority traffic MUST NOT be given preference over higher priority traffic. It is RECOMMENDED that at least two TCP connections be provided, for example one for priority 6 and 7 traffic and one for priority 0 through 5 traffic, in order to insure that urgent control traffic, including control traffic related to establishing and maintaining adjacency, is not significantly delayed by lower priority traffic.

TCP is a stream protocol, not a record oriented protocol, so a TRILL data packet with its header or a TRILL IS-IS packet might be split across multiple TCP packet payloads or a single TCP packet payload could include multiple TRILL packets or the like. Thus a framing mechanism is needed, as specified below, so that a received TRILL stream can be parsed into TRILL packets.

In the TCP header, the source and destination port fields are as follows:

Source Port - along with Source IP, Destination IP, and

Destination Port, identifies a TCP flow.

Destination Port - indicates TRILL Data or IS-IS, see Section 11.1.

TRILL packets are framed for transmission over TCP as shown below.

```
+-----+----- // ----+
| Length | TRILL packet |
+-----+----- // ----+
```

Length - the length of the TRILL packet in bytes as a 2-byte unsigned integer in network order.

TRILL packet - The TRILL packet within framing starts with the TRILL or the L2-IS-IS Ethertype (0x22F3 or 0x22F4). If the initial 2 bytes of the TRILL packet are not the correct Ethertype based on the Destination Port, then the receiver assumes that framing synchronization has been lost and MUST close that TCP connection. Note that the Hamming distance between these Ethernets is 2 so that a single bit error cannot convert one into the other.

The sequence of framed TRILL packets is sliced as necessary into TCP packet payloads.

Depending on performance requirements, in many cases consideration should be given to tuning TCP. Methods for doing this are out of scope for this document. See [RFC7323].

5.6.1 TCP Connection Establishment

If a TRILL over IP transport port is configured to always use TCP it will also be configured with a list of IP addresses and MUST try to establish a TCP connection to each of them. It also MUST accept TCP connections from each of that list of IP addresses.

If a TRILL over IP port supports TCP but is using UDP for neighbor discovery and encapsulation negotiation, then it MUST try to establish a TCP connection to any adjacent port in the Report state (see [RFC7177] and Section 5.2) when TCP has been negotiated with that port. It also MUST accept TCP connections from each such adjacent port.

Establishing a connection actually means to initiate TCP connections for each DSCP value that the TRILL over IP port is configured to use in TCP communication with the destination separately for TRILL Data and TRILL IS-IS as they have different destination ports, unless such

a connection already exists. For example, port P1 could meet the requirements to establish a TCP connection to port P2 and find that such a connection already exists having been initiated by P2. A TCP connection can be used bi-directionally for TRILL traffic. However the timing and implementation details may be such that P1 and P2 each establish a TCP connection to the other, in which case it might be that each of those connections would be used uni-directionally for TRILL traffic.

When a TCP connection is closed or reset, if the conditions are still met for that TCP port to establish that connection, it waits a configurable length of time that defaults to 80 milliseconds and tried to re-establish the connection. See Section 9.2.3.3.

5.7 Other Encapsulations

Additional TRILL over IP transport encapsulations may be specified in future documents and allocated a link technology specific flag bit as per Section 11.3. A primary consideration for whether it is worth the effort to specify use of an encapsulation by TRILL over IP transport is whether it has good existing or anticipated fast path support.

6. Handling Multicast

For UDP based encapsulations where IPsec is not in use, both IS-IS packets and multi-destination TRILL Data packets are, by default, sent to an All-RBridges IPv4 or IPv6 IP multicast address as appropriate (see Section 11.2); however, such a TRILL over IP transport 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 transport ports to which multi-destination packets are sent. In the serial unicast case the outer IP header of each copy of the a TRILL Data packet sent shows an IP unicast destination address even though, for TRILL Data packets, the TRILL header has the M bit set to one to indicate multi-destination.

Serial unicast configuration MUST be used if the TRILL over IP transport port (1) is connected to an IP network does not support IP multicast, (2) uses TCP based encapsulation, or (3) is using IPsec. In any case, unicast TRILL Data packets (those with the M bit in the TRILL Header set to zero) are sent by unicast IP. If a TRILL over IP transport port is configured to send all traffic secured and/or with TCP, adjacency and data flow will only be possible with IP addresses in a configured list at that port (see Section 9).

Even if a TRILL over IP transport port is configured to send multi-destination packets with serial unicast, if it uses a UDP based encapsulation it MUST be prepared to receive IP multicast TRILL packets. TRILL over IP transport ports that are using multicast default to periodically transmitting appropriate IGMP (IPv4 [RFC3376]) or MLD (IPv6 [RFC2710]) packets, so that the TRILL multicast IP traffic can be sent to them, but MAY be configured not to do so.

There may be good reasons for configuring TRILL over IP transport ports to use serial unicast even where native IP multicast is available and could be used. Use of serial unicast provides the network manager with more precise control over adjacencies and how TRILL over IP transport 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 transport connections between two 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 ports that support TRILL over IP transport MUST implement IPsec [RFC4301] and support the use of IPsec Encapsulating Security Protocol (ESP [RFC4303]) in tunnel mode to secure both IS-IS and TRILL Data packets. When IPsec is used to secure a TRILL over IP transport link and no IS-IS security is enabled, the IPsec session MUST be fully established before any IS-IS or TRILL data packets are exchanged. When there is IS-IS security [RFC5310] provided, implementers SHOULD use IS-IS security to protect IS-IS packets. However, in this case, the IPsec session still MUST be fully established before any TRILL Data packets transmission, since IS-IS security does not provide any protection for data packets, and the IPsec session SHOULD be fully established before any IS-IS packet transmission other than IS-IS Hello or MTU PDUs.

All RBridges that support TRILL over IP transport 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 SHOULD use a pre-shared key that incorporates the IS-IS shared key. 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. P1-System-ID and P2-System ID are the six byte System IDs of the two TRILL RBridges, and P1-Port and P2-Port are the TRILL Port IDs [RFC6325] of 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. Note that the value to which the HKDF function is applied starts with 0x54 (the

ASCII code for "T") while the data to which [RFC5310] authentication is applied (an IS-IS PDU) starts with 0x83, the Interdomain Routing Discriminator, thus, although they are both SHA256 based, they are never applied to the same value.

With [RFC5310] there could be multiple keys identified with 16-bit key IDs. The key ID when an IS-IS key is in use is transmitted in an IKEv2 ID_KEY_ID identity field [RFC7296] with Identification Data length of 2 bytes (Payload Length 6 bytes). The Key ID of the IS-IS-key is used to identify the IKEv2 shared secret derived as above that is actually used. ID_KEY_ID identity field(s) of other lengths MAY occur but their use is beyond the scope of this document.

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 an IS-IS shared key MUST expire within a lifetime no longer than the IS-IS-shared key from which they were derived. When the IKEv2 shared secret key expires, or earlier, the IKEv2 Security Association must be rekeyed using a new shared secret derived from a new IS-IS shared key.

IKEv2 with certificate-based security MAY be used but details of certificate contents and use policy for this application of IKEv2 are beyond the scope of this document.

7.1.2 Group Keying

In the case of a TRILL over IP transport port configured as point-to-point (see Section 4.2.4.1 of [RFC6325]), there is no group keying and the pairwise keying determined as provided in Section 7.1.1 is used for multi-destination TRILL traffic, which is unicast across the link.

In the case of a TRILL over IP transport 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 multi-destination TRILL traffic, which is unicast across the link.

The case of a TRILL over IP transport port configured as broadcast and using native multicast is beyond the scope of this document and is expected to be covered in a future document [SGKPuses]. For security as provided in this document, multicast is handled via serial unicast.

7.2 Mandatory-to-Implement Algorithms

All RBridges that support TRILL over IP transport MUST implement IPsec ESP [RFC4303] in tunnel mode. The implementation requirements for ESP cryptographic algorithms are as specified for IPsec. That specification is currently [RFC8221].

8. Transport Considerations

This section discusses a variety of important transport considerations. NAT traversal is out of scope for this document.

8.1 UDP Congestion Considerations

This subsection discusses TRILL over UDP congestion considerations. These are applicable to the UDP based TRILL over IP transport encapsulation headers specified in detail in this document. Other encapsulations would likely have different congestion considerations and, in particular, the TCP encapsulation specified in Section 5.6 does not need congestion control beyond that provided by TCP. Congestion considerations for additional TRILL encapsulations will be provided in the document specifying that encapsulation.

One motivation for including UDP or TCP as the outermost part of a TRILL over IP encapsulation header is to improve the use of multipath such as Equal Cost Multi-Path (ECMP) in cases where traffic is to traverse routers that are able to hash on Port and IP address through addition of entropy in the source port (see Section 8.3). 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.

Section 3.1.11 of [RFC8085] discusses the congestion considerations for design and use of UDP tunnels; this is important because other flows could share the path with one or more UDP tunnels, necessitating congestion control [RFC2914] to avoid destructive interference.

The default initial determination of the TRILL over IP transport encapsulation to be used is through the exchange of IS-IS Hellos. This is a low bandwidth process. Hellos are not permitted to be sent any more often than once per second, and so are very unlikely to cause congestion. Thus no additional controls are needed for Hellos even if they are sent, as is the default, over UDP.

Congestion has potential impacts both on the rest of the network containing a UDP flow and on the traffic flows using the UDP encapsulation. These impacts depend upon what sort of traffic is carried in UDP, as well as the path it follows. The UDP based TRILL over IP transport encapsulations specified in this document do not provide any congestion control and are transmitted as regular UDP packets.

The use of serial unicast, where the transmission of a multi-destination TRILL packet is executed as multiple unicast transmission, potentially increases link load and could thus increase congestion. Rate limiting of multi-destination traffic that is to be transmitted in this fashion should be considered.

The subsections below discuss congestion for TRILL over IP transport traffic with UDP based encapsulation headers in traffic-managed controlled environments (TMCE, see [RFC8086]) and other environments.

8.1.1 Within a TMCE

Within a TMCE, that is, an IP network that is traffic-engineered and/or otherwise managed, for example via use of traffic rate limiters, to avoid congestion, UDP based TRILL over IP encapsulation headers are appropriate for carrying traffic that is not known to be congestion controlled. In such cases, operators of TMCE networks avoid congestion by careful provisioning of their networks, rate-limiting of user data traffic, and traffic engineering according to path capacity.

When TRILL over IP transport using a UDP based encapsulation header carries traffic that is not known to be congestion controlled in a TMCE network, the traffic path **MUST** be entirely within that network, and measures **SHOULD** be taken to prevent the traffic from "escaping" the network to the general Internet. Examples of such measures are:

- o physical or logical isolation of the links carrying the traffic from the general Internet and
- o deployment of packet filters that block the UDP ports assigned for TRILL over IP transport.

8.1.2 In Other Environments

Where UDP based encapsulation headers are used in TRILL over IP transport in environments other than those discussed in Section 8.1.1, specific congestion control mechanisms such as rate limiting are commonly needed. However, if the traffic being carried by the TRILL over IP transport link is already congestion controlled and the size and volatility of the IS-IS link state database is limited, then specific congestion control may not be needed. See [RFC8085] Section 3.1.11 for further guidance.

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 a TRILL(EF) packet that was subsequently sent to a next hop RBridge out a TRILL over IP transport 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 around, being further re-ingressed, and so on. (Or this might occur through a cycle of TRILL different campuses.) The packet would get discarded if it got too large unless fragmentation is enabled, in which case 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 transport 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 transport 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 TRILL packets over any Equal Cost Multiple Paths (ECMPs) existing internal to the IP path between TRILL over IP transport ports.

The ECMP election for the IP traffic could be based, for example with 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 UDP Source Port as an entropy field. (This idea is also introduced in [RFC8086].) 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.Label. These are fields from the TRILL data payload which looks like an Ethernet frame (see [RFC7172] Figures 1 and 2).

8.4 MTU Considerations

In TRILL each RBridge 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 cannot support 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 the largest LSPs.

A method of determining `originatingLSPBufferSize` for an RBridge is described in [RFC7780]. If that RBridge has a TRILL over IP transport port that either (1) can accommodate jumbo frames, (2) is a link on which IP fragmentation is enabled and acceptable, or (3) is configured to use TCP encapsulation for all packets, then it is unlikely that the port will be a constraint on the `originatingLSPBufferSize` of the RBridge. On the other hand, if the TRILL over port can only handle smaller frames, a UDP encapsulation is in use at least for Hellos, and fragmentation is to be avoided when possible, a TRILL over IP transport port might have an MTU that constrained the RBridge's `originatingLSPBufferSize`.

Because TRILL sets the minimum value of *Sz* at 1,470 bytes, R Bridges will not constrain LSPs or other IS-IS PDUs to a size smaller than that. Therefore there may be TRILL over IP transport ports that require that either fragmentation be enabled or that TCP based encapsulation for all TRILL packets be used if TRILL communication over that IP port is desired. When fragmentation is enabled or TCP is in use, the effective link MTU from the TRILL point of view is larger than the RBridge port to RBridge port path MTU from the IP point of view.

TRILL IS-IS MTU PDUs, as specified in Section 5 of [RFC6325] and in [RFC7177], MUST NOT be fragmented when sent over UDP and can be used to obtain added assurance of the MTU of a link. The algorithm discussed in [RFC8249] should be useful in determining the IP MTU between a pair of RBridge ports that have IP connectivity with each other. See also [RFC4821].

An appropriate time to confirm MTU, or re-discover it if it has changed, is when an RBridge notices topology changes in a path between RBridge ports that is in use for TRILL over IP transport; however, MTU can change at other times. For example, if two RBridge ports are connected by a bridged LAN, topology or configuration changes within that bridged LAN could change the MTU between those RBridge ports.

For further discussion of these issues, see [IntareaTunnels].

9. TRILL over IP Transport Port Configuration

This section specifies the configuration information needed at a TRILL over IP transport port beyond that needed for a general RBridge port.

9.1 Per IP Port Configuration

Each RBridge port used for a TRILL over IP transport 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 transport 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 transport port IP address.

The mapping from TRILL packet priority to TRILL over IP transport Differentiated Services Code Point (DSCP [RFC2474]) can be configured. If supported, mapping from an inner DSCP code point, when the TRILL payload is IP, to the outer TRILL over IP transport DSCP can be configured. (See Section 4.3.)

Each TRILL over IP transport port has a list of acceptable encapsulations it will use as the basis of adjacency. By default this list consists of one entry for native encapsulation (see Section 7). Additional encapsulations MAY be configured and native encapsulation MAY be removed from this list by configuration. 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 transport 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 transport port is configured whether or not to use IPsec (Section 9.2.4).

Regardless of whether they will send IP multicast, TRILL over IP transport ports emit appropriate IGMP (IPv4 [RFC3376]) or MLD (IPv6 [RFC2710]) packets unless configured not to do so. These are sent for the IP multicast group the port would use if it sent IP multicast.

9.2.1 Native Multicast Configuration

If a TRILL over IP transport port address is using native IP multicast for multi-destination packets (IS-IS and TRILL 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 transport port may be configured to use a different IP multicast address for multicasting packets.

9.2.2 Serial Unicast Configuration

If a TRILL over IP transport port address has been configured to use serial unicast for multi-destination packets (IS-IS and TRILL data), it has 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 over IP packets are serially unicast to the addresses in this list. Such a TRILL over IP transport 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 IS-IS Hellos. IP packets received from a source IP address not on the list are discarded.

If this list of destination IP addresses is empty, the port is disabled.

9.2.3 Encapsulation Specific Configuration

Specific TRILL over IP transport encapsulation methods may provide for further configuration as specified below.

9.2.3.1 UDP Source Port

As discussed above, the UDP based encapsulations (Sections 5.4 and 5.5) start with a header containing a source port number that can be used for entropy (Section 8.3). The range of source port values used defaults to the ephemeral port range (49152-65535) but can be configured to any other range.

9.2.3.2 VXLAN Configuration

A TRILL over IP transport port using VXLAN encapsulation can be configured with non-default VXLAN Network Identifiers (VNIs) that are used in that field of the VXLAN header for all IS-IS and TRILL Data packets sent using the encapsulation and required in those received using the encapsulation. The default VNI is 1 for IS-IS and 2 for TRILL Data. A TRILL packet received with the an unknown VNI is discarded.

A TRILL over IP transport port using VXLAN encapsulation can also be configured to map the Inner.VLAN of a TRILL Data packet being transported to the value it places in the VNI field and/or to copy or map the Inner.FGL [RFC7172] of a TRILL Data packet to the VNI field.

9.2.3.3 TCP Configuration

A TRILL over IP transport port using TCP encapsulation is configurable as to the connection re-establishment delay in the range of 1 to 10,000 milliseconds that defaults to 80 milliseconds. See Section 5.6.1.

9.2.3.4 Other Encapsulation Configuration

Additional encapsulation methods, beyond those specified in this document, are expected to be specified in future documents and may require further configuration.

9.2.4 Security Configuration

A TRILL over IP transport port can be configured, for the case where IS-IS security [RFC5310] is in use, as to whether or not IPsec must be fully established and used for any IS-IS transmissions other than IS-IS Hello or MTU PDUs (see Section 7). There may also be configuration whose details are outside the scope of this document concerning certificate based IPsec or use of shared keys other than IS-IS based shared key or how to select the IS-IS based shared key to use.

10. Security Considerations

TRILL over IP transport 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 3 above.

For communication between end stations in a TRILL campus, security may be possible at three levels: end-to-end security between those end stations, edge-to-edge security between ingress and egress RBridges, and link security to protect a TRILL hop. Any combination of these can be used, including all three.

TRILL over IP transport link security protects the contents of TRILL Data and IS-IS packets over a single TRILL hop between RBridge ports, including protecting the identities of the end stations for data and the identities of the edge RBridges, from observers of the link and transit devices within the link such as bridges or IP routers, but does not encrypt the link local IP addresses used in a packet and does not protect against observation by the RBridges on the link.

Edge-to-edge TRILL security would protect the contents of TRILL data packets between the ingress and egress RBridges, including the identities of the end stations for data, from transit RBridges but does not encrypt the identities of the edge RBridges involved and does not protect against observation by those edge RBridges. Edge-to-edge TRILL security may be covered in future documents.

End-to-end security does not protect the identities of the end stations or edge RBridge involved but does protect the user data content of TRILL data packets from observation by all RBridges or other intervening devices between the end stations involved. End-to-end security should always be considered as an added layer of security to protect any particularly sensitive information from unintended disclosure. Such end-station to end-station security is generally outside the scope of TRILL

If VXLAN encapsulation is used, the unused Ethernet source and destination MAC addresses mentioned in Section 5.5, provide a 96 bit per packet side channel.

10.1 IPsec

This document specifies that all RBridges that support TRILL over IP transport 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 transport link will traverse a network that is not under the same administrative control as the rest of the TRILL campus or is not 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 transport 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 transport 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 Ports in the Service Name and Transport Protocol Port Number Registry [PortRegistry] for TRILL Data and L2-IS-IS as shown below. The L2-IS-IS port is used for the transmission of IS-IS PDUs by TRILL and may be used other protocols.

It is requested that the Hamming distance between the two port number be at least 2, that is, that at least two bits differ between the port numbers. For example, they could be an odd number and the following even number such that both of the bottom two bits would differ between them.

Service Name: TRILL-data
Transport Protocol: udp, tcp
Assignee: iesg@ietf.org
Contact: chair@ietf.org
Description: Transport of TRILL Data packets.
Reference: [this document]
Port Number: (TBD2)

Service Name: L2-IS-IS
Transport Protocol: udp, tcp
Assignee: iesg@ietf.org
Contact: chair@ietf.org
Description: Transport of IS-IS PDUs.
Reference: [this document]
Port Number: (TBD1)

11.2 Multicast Address Assignments

IANA is requested to assign one IPv4 and one IPv6 multicast address, as shown below, which correspond to both the All-RBridges and All-IS-IS-RBridges multicast MAC addresses that have been assigned for TRILL. Because the low level hardware MAC address dispatch considerations for TRILL over Ethernet do not apply to TRILL over IP transport, one IP multicast address for each version of IP is sufficient.

(Value recommended to IANA in square brackets)

Name	IPv4	IPv6
-----	-----	-----
All-RBridges	TBD3	TBD4[FF0X:::15D]

The hex digit "X" in the IPv6 variable scope 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	
0x007-0xFF7	Unassigned
NEW	
0x007-0xFCF	Unassigned
0xFD0-0xFF7	(link technology dependent, see subregistry)

A new indented subregistry under the re-named "RBridge Channel Protocols and Link Technology Specific Flags" registry is added as follows:

Name: TRILL over IP Transport Link Flags
Registration Procedure: Expert Review
Reference: [this document]

Flag	Meaning	Reference
-----	-----	-----
0xFD0	Native encapsulation fully supported	[this document]
0xFD1	VXLAN encapsulation fully supported	[this document]
0xFD2	TCP encapsulation fully supported	[this document]
0xFD3-0xFF7	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>>.
- [RFC0793] - Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<http://www.rfc-editor.org/info/rfc793>>.
- [RFC1122] - Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <<https://www.rfc-editor.org/info/rfc1122>>.
- [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>>.
- [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 (RBriges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <<http://www.rfc-editor.org/info/rfc6325>>.
- [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, DOI 10.17487/RFC7175, May 2014, <<http://www.rfc-editor.org/info/rfc7175>>.
- [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>>.
- [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>>.

- [RFC7780] - Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <<http://www.rfc-editor.org/info/rfc7780>>.
- [RFC8174] - Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8200] - Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8221] - Wouters, P., Migault, D., Mattsson, J., Nir, Y., and T. Kivinen, "Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 8221, DOI 10.17487/RFC8221, October 2017, <<https://www.rfc-editor.org/info/rfc8221>>.
- [RFC8249] - Zhang, M., Zhang, X., Eastlake 3rd, D., Perlman, R., and S. Chatterjee, "Transparent Interconnection of Lots of Links (TRILL): MTU Negotiation", RFC 8249, DOI 10.17487/RFC8249, September 2017, <<https://www.rfc-editor.org/info/rfc8249>>.
- [LEphb] - R. Bless, "A Lower Effort Per-Hop Behavior)LE PHB)", draft-ietf-tsvwg-le-phb, work in progress.

Informative References

- [RFC4821] - Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <<http://www.rfc-editor.org/info/rfc4821>>.
- [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>>.
- [RFC7323] - Borman, D., Braden, B., Jacobson, V., and R. Scheffenegger, Ed., "TCP Extensions for High Performance", RFC 7323, DOI 10.17487/RFC7323, September 2014, <<https://www.rfc-editor.org/info/rfc7323>>.
- [RFC7657] - Black, D., Ed., and P. Jones, "Differentiated Services (Diffserv) and Real-Time Communication", RFC 7657, DOI 10.17487/RFC7657, November 2015, <<https://www.rfc-editor.org/info/rfc7657>>.
- [RFC8085] - Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <<http://www.rfc-editor.org/info/rfc8085>>.
- [RFC8086] - Yong, L., Ed., Crabbe, E., Xu, X., and T. Herbert, "GRE-in-UDP Encapsulation", RFC 8086, DOI 10.17487/RFC8086, March 2017, <<http://www.rfc-editor.org/info/rfc8086>>.
- [IntareaTunnels] - J. Touch, M. Townsley, "IP Tunnels in the Internet Architecture", draft-ietf-intarea-tunnels, work in progress.
- [TRILLECN] - Eastlake, D., B. Briscoe, "TRILL: ECN (Explicit Congestion Notification) Support", draft-ietf-trill-ecn-support, work in progress.
- [SGKPuses] - D. Eastlake, D. Zhang, "Simple Group Keying Protocol TRILL Use Profiles", draft-ietf-trill-link-gk-profiles, work in progress.

[PortRegistry] - <https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml>

Appendix A: IP Security Choice

This informational appendix discusses the choice of mandatory to implement IP security protocol for TRILL over IP transport ports. Other security protocols can be used by agreeing TRILL over IP transport ports, but one protocol was selected as mandatory to implement for interoperability.

The TRILL WG considered both DTLS and IPsec as the mandatory to implement IP security protocol. Perhaps the most extensive discussion occurred at the TRILL WG meeting at IETF meeting 91. The WG decided to go with IPsec due to better hardware support which was considered an important factor for being able to operate at or near line speed. Tunnel mode was chosen as there appeared to be better support for it in offboard hardware devices.

Acknowledgements

The following people have provided useful feedback on the contents of this document: Sam Hartman, Adrian Farrel, Radia Perlman, Ines Robles, Mohammed Umair, Magnus Westerlund, and Lucy Yong.

Some of the material in this document is derived from [RFC8085] and [RFC8086].

Authors' Addresses

Margaret Cullen
Painless Security
14 Summer Street, Suite 202
Malden, MA 02148
USA

Phone: +1-781-605-3459
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
Huawei Technologies

Email: dacheng.zhang@huawei.com

Copyright, Disclaimer, and Additional IPR Provisions

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

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

TRILL WG
INTERNET-DRAFT
Intended Status: Informational
Expires: October 20, 2017

R. Parameswaran,
Brocade Communications, Inc.
April 22, 2017

TRILL: Parent node Shifts in Tree Construction, Mitigation.
<draft-rp-trill-parent-selection-03.txt>

Abstract

This draft documents a known problem in the TRILL tree construction mechanism and offers an approach requiring no change to the TRILL protocol in order to solve the problem.

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>.

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].

Table of Contents

1. Introduction.....	1
2. Tree construction in TRILL.....	2
3. Issues with the TRILL tree construction algorithm.....	2
4. Solution using the Affinity sub-TLV.....	4
5. Network wide selection of computation algorithm.....	7
6. Relationship to draft-ietf-trill-resilient-trees.....	7
7. Security Considerations.....	9
8. IANA Considerations.....	9
9. Informative References.....	9

1. Introduction.

TRILL is a data center technology that uses link-state routing mechanisms in a layer 2 setting, and serves as a replacement for spanning-tree. TRILL uses trees rooted at pre-determined nodes as a way to distribute multi-destination traffic. Multi-destination traffic includes traffic such as layer-2 broadcast frames, unknown unicast flood frames, and layer 2 traffic with multicast MAC addresses (collectively referred to as BUM traffic). Multi-destination traffic is typically hashed onto one of the available trees and sent over the tree, potentially reaching all nodes in the network (hosts behind which may own/need the packet in question).

2. Tree construction in TRILL.

Tree construction in TRILL is defined by [RFC6325], with additional corrections defined in [RFC7780].

The tree construction mechanism used in TRILL codifies certain tree construction steps which make the resultant trees very brittle. Specifically, the parent selection mechanism in TRILL causes problems in case of node failures. TRILL uses the following rule - when constructing an SPF tree, if there are multiple possible parents for a given node (i.e. if multiple upstream nodes can potentially pull in a given node during SPF, all at the same cumulative cost, then the parent selection is imposed in the following manner):

[RFC6325]:

"When building the tree number j , remember all possible equal cost parents for node N . After calculating the entire 'tree' (actually, directed graph), for each node N , if N has ' p ' parents, then order the parents in ascending order according to the 7-octet IS-IS ID considered as an unsigned integer, and number them starting at zero. For tree j , choose N 's parent as choice $j \bmod p$."

There is an additional correction posted to this in [RFC7780]:

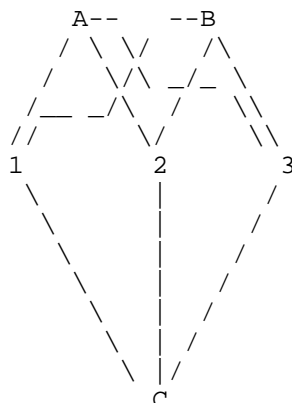
[RFC7780], Section 3.4:

"Section 4.5.1 of [RFC6325] specifies that, when building distribution tree number j , node (RBridge) N that has multiple possible parents in the tree is attached to possible parent number $j \bmod p$. Trees are numbered starting with 1, but possible parents are numbered starting with 0. As a result, if there are two trees and two possible parents, then in tree 1 parent 1 will be selected, and in tree 2 parent 0 will be selected.

This is changed so that the selected parent MUST be $(j-1) \bmod p$. As a result, in the case above, tree 1 will select parent 0, and tree 2 will select parent 1. This change is not backward compatible with [RFC6325]. If all RBridges in a campus do not determine distribution trees in the same way, then for most topologies, the RPFC will drop many multi-destination packets before they have been properly delivered."

3. Issues with the TRILL tree construction algorithm.

With this tree construction mechanism in mind, let's look at the Spine-Leaf topology presented below and consider the calculation of Tree number 2 in TRILL. Assume all the links in the tree are at the same cost.



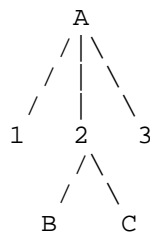
Assume that in the above topology, when ordered by 7-octet ISIS-id,

$1 < 2 < 3$ holds and that the root for Tree number 2 is A. Given the ordered set $\{1, 2, 3\}$, these nodes have the following indices (with a starting index of 0):

Node	Index
1	0
2	1
3	2

Given the SPF constraint and that the tree root is A, the parent for nodes 1, 2, and 3 will be A. However, when the SPF algorithm tries to pull B or C into the tree, we have a choice of parents, namely 1, 2, or 3.

Given that this is tree 2, the parent will be the one with index $(2-1) \bmod 3$ (which is equal to 1). Hence the parent for node B will be node 2.

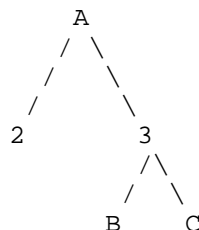


However, due to TRILL's parent selection algorithm, the sub-tree rooted at Node 2 will be impacted even if Node 1 or Node 3 go down.

Take the case where Node 1 goes down. Tree 2 must now be re-computed (this is normal) - but now, when the SPF computation is underway, when the SPF process tries to pull in B, the list of potential parents for B now are $\{2 \text{ and } 3\}$. So, after ordering these by ISIS-Id as $\{2, 3\}$ (where 2 is considered to be at index of 0 and 3 is considered to be at index 1), for tree 1, we apply TRILL's formula of:

Parent's index = $(\text{TreeNumber}-1) \bmod \text{Number_of_parents}$.
 $= (2-1) \bmod 2$
 $= 1 \bmod 2$
 $= 1$ (which is the index of Node 3)

The re-calculated tree now looks as shown below. The shift in parent nodes (for B) may cause disruption to live traffic in the network, and is unnecessary in absolute terms because the existing parent for node B, node 2, was not perturbed in any way.



Aside from the disruption posed by the change in the tree links, depending upon how the concerned rbridges stripe vlans/FGLs across trees and how they may prune these, additional disruption is possible if the forwarding state on the new parent rbridge is not primed to match the new tree structure. This churn could simply be avoided with a better approach.

The parent shift issue noted above can be solved by using

the Affinity sub-TLV.

While the technique identified in this draft has an immediate benefit when applied to spine/leaf networks popular in data-center designs, nothing in the approach outlined below assumes a spine-leaf network. The technique presented below will work on any connected graph. Furthermore, no directional symmetry in link-cost is assumed.

4. Solution using the Affinity sub-TLV.

At a high level, this problem can be solved by having the affected parent send out an Affinity sub-TLV identifying the children for which it wants to preserve the parent-child relationship, subject to network events which may change the structure of the tree. The affected parent node would send out an Affinity sub-TLV with multiple Affinity records, one per child node, listing the concerned tree number.

It would be sufficient to have a local configuration option (e.g. a CLI) at one of the nodes which is deemed to be the parent of choice (referred to as designated parent below). The following steps provide a way to implement this proposal:

- a. The operator locally configures the designated parent to indicate its stickiness in tree construction for a specific tree number and tree root via the Affinity sub-TLV. This can be done before tree construction if the operator consults the 7 octet ISIS-ID relative ordering of the concerned nodes and decides up-front which of the potential parent nodes should become the parent node for a given set of children on that tree number under the TRILL tree construction mechanism. The operator **MUST** configure the designated parent stickiness on only one node amongst a set of sibling (potential parent) nodes relative to the tree root for that tree number. It is suggested that the parent stickiness be configured on the node that would have been selected as the parent under default Trill parent selection rules. Parent stickiness **MUST NOT** be configured on the root of the tree, or if configured previously on a non-root node with the root for that tree shifting to that node subsequently, such configuration **MUST** be ignored on the root node.
- b. On any subsequent SPF calculation after the operator configures the designated parent as indicated above, when the designated parent node finds that it could be a potential parent for one or more child nodes during tree construction, it declares itself to be the parent for the concerned child nodes, over-riding the default TRILL parent selection rules. The configured node advertises its parent preference via the Affinity sub-TLV when it completes a tree calculation, and finds itself the parent of one or more child nodes per the SPF tree calculation. The Affinity sub-TLV **MUST** reflect the appropriate tree number and the child nodes for which the concerned node is a parent node. The Affinity sub-TLV **SHOULD** be published when the tree computation is deemed to have converged (more on this under d. below).
- c. Likewise, when any change event happens in the network, one which forces a tree re-calculation for the concerned tree, the designated parent node should run through the normal TRILL tree calculation agnostic of the fact that it has published an Affinity sub-TLV as well as agnostic of the default TRILL tree selection rules i.e the node asserts its right to be a parent without directly referencing either the default Trill parent selection rules or its own published Affinity sub-TLV in establishing parent relationships.
- d. During the SPF tree calculation, the designated parent node should react in the following manner:

- i. If the node is a potential parent for some of the children identified in an existing Affinity sub-TLV, if any, after convergence of the tree computation, the node MUST send out an (updated) Affinity sub-TLV identifying the correct sub-set of children for which the node aspires to establish/continue the parent relationship. This case would also apply if there are new child nodes for which the node is now a parent (however, see the conflicted Affinity sub-TLV rules in vii and j. below).

For its own tree computation, the designated parent node MUST use itself as parent in order to pull the set of children identified during the SPF run into the tree, barring a conflicting affinity sub-TLV seen from another node (see vii. below for handling this case).

- ii. If the tree structure changes such that the designated node is no longer a potential parent for any of the child nodes in the advertised Affinity sub-TLV, then it SHOULD retract the Affinity sub-TLV, upon convergence of the tree computation. In this case, the default TRILL tie-break rule would need to be used during SPF construction for the nodes that were children of this designated node previously. One specific case may be worth high-lighting - if a parent-child relationship inverts i.e. if the designated parent becomes a child of its former child node due to a change in the tree structure, it MUST exclude that child from its Affinity sub-TLV. In such case, if the designated parent node cannot maintain a parent relationship with any of its prior child nodes, then it MUST retract any previously published affinity sub-TLV.
- iii. Nodes SHOULD use a convergence timer to track completion of the tree computation. If there are any additional tree computations while the convergence timer is running, the timer SHOULD be re-started/extended in order to absorb the interim network events. It is possible that the intended action at the expiration of the timer may change meanwhile. The timer needs to be large enough to absorb multiple network events that may happen due to a change in the physical state of the network, and yet short enough to avoid delaying the update of the Affinity sub-TLV.
- iv. At the expiration of the convergence timer, the existing state of the tree MUST be compared with the existing Affinity sub-TLV and the intended change in the status of the Affinity sub-TLV is carried out e.g. a fresh publication, or an update to the list of children, or a retraction.
- v. Alternately, the above steps (re-examination of the Affinity sub-TLV and update) MAY be tied to/triggered from the download of the tree routes to the L2 RIB, since that typically happens upon a successful computation of the complete tree. An additional stabilization timer could be used to counteract back-to-back L2 RIB downloads due to repeated computations of the tree due to a burst of network events.
- vi. Note that this approach may cause an additional tree computation at remote nodes once the updated Affinity sub-TLV (or lack of it) is received/perceived, beyond the network events which led up to the change in the tree. In the case where an operator introduced a designated parent configuration on an existing tree, then remote nodes would need to receive the Affinity sub-TLV indicating the designated parent's Affinity for its children before the remote nodes shift away from the default TRILL parent selection rules. However, in most cases, in steady state, this mechanism should cause very little tree churn unless

a designated parent configuration was introduced, removed, or a link between the designated parent and its children changed state. In cases where the network change event originated on the designated parent node, it may be possible to optimize on the churn by packing both the data bearing the network change event and the Affinity sub-TLV into the same link-state update packet.

- vii. In situations where the designated parent node would normally originate an affinity sub-TLV to indicate affinity to a specific set of child nodes, it MUST NOT originate an Affinity sub-TLV if it sees an Affinity sub-TLV from some other node for the same tree number and for all of the same child-nodes, such that the other node's Affinity sub-TLV would win using the conflict tie-break rules in section 5.3 of [RFC7783]. Any existing Affinity sub-TLV already published by this node in such a situation MUST be retracted. If only some of the child nodes overlap between the two conflicting Affinity sub-TLVs, then this designated parent node MAY continue to publish its affinity sub-TLV listing its child nodes that are not in conflict with the other Affinity sub-TLV. Other guide-lines listed in [RFC7783] MUST be adhered to as well - the originator of the Affinity sub-TLV must name only directly adjacent nodes as children, and must not name the tree root as a child.
- e. Situations where the node advertising the Affinity sub-TLV dies or restarts SHOULD be handled using the normal handling for such scenarios relating to the parent Router Capability TLV, and as specified in [RFC4971].
- f. Situations where a parent-child link directly connected to the designated parent node constantly flaps, MUST be handled by having the designated parent node retract the Affinity sub-TLV, if it affects the parent-child relationships in consideration. The long-term state of the Affinity sub-TLV can be monitored by the designated parent node to see if it is being published and retracted repeatedly in multiple iterations or if a specific set of children are being constantly added and removed. The designated parent may resume publication of the Affinity sub-TLV once it perceives the network to be stable again in the future.
- g. If the designated parent node is forced to retract its Affinity sub-TLV due to a change in the tree structure, it can then repeat these steps in a subsequent tree construction, if the same node becomes a parent again, so long as it perceives its parent-child links to be stable (free of link/node flaps).
- h. In terms of nodes that do not support this draft, they are expected to seamlessly inter-operate with this draft, so long as they understand and honor the Affinity sub-TLV. The draft assumes that most TRILL implementations now support the Affinity sub-TLV. In any case, the guide-lines specified in section 4.1 of [RFC7783] MUST be used i.e. if all nodes in the network do not support the Affinity sub-TLV then the network must default to the Trill parent selection rules.
- i. Remote nodes MUST default to the Trill parent selection rules if they do not see an Affinity sub-TLV sent by any node in the network.
- j. At remote nodes, conflicting Affinity sub-TLVs from different originators for the same tree number and child node MUST be handled as specified in section 5.3 of [RFC7783], namely by selecting the Affinity sub-TLV originated by the node with the highest priority to be a tree root, with System-ID as tie-breaker.

5. Network wide selection of computation algorithm.

The proposed solution above does not need any operational change to the TRILL protocol, beyond the usage of the Affinity sub-TLV (which is already in the proposed standard) for the use case identified in this draft.

6. Relationship to draft-ietf-trill-resilient-trees.

Given that both draft-ietf-trill-resilient-trees, and draft-rp-trill-parent-selection-03 drafts use the Affinity sub-TLV, it is worthwhile to examine if there is any functional overlap between the two drafts. At a high level, the two drafts have different goals and appear to solve unrelated problems.

draft-ietf-trill-resilient-trees relates to link protection, and defines the notion of a primary distribution tree and a backup distribution tree (DT), where these trees are intentionally kept link disjoint to the extent possible, and the backup tree is pre-programmed in the hardware, and activated either up front or upon failure of the primary distribution tree.

On the other hand, draft-rp-trill-parent-selection-03 protects parent-child relationships of interest on the primary DT, and has no direct notion of a backup DT.

draft-ietf-trill-resilient-trees considers the following algorithmic approaches to the building the backup distribution tree (section numbers listed below are from draft-ietf-trill-resilient-trees):

1. Operator hand-configuration for links on the backup DT/manual generation of Affinity sub-TLV - this is very tedious and unlikely to scale or be implemented in practice, and hence is disregarded in the analysis here.
2. Section 3.2.1.1a: Use of MRT algorithms (which will produce conjugate trees - link disjoint trees with roots for primary and backup trees that are coincident on the same rBridge).
3. Section 3.2.1.1b: Once the primary DT is constructed, the links used in the primary DT are additively cost re-weighted, and a second SPF is run to derive the links comprising the backup DT. Affinity sub-TLV is used to mark links on the back-up DT which are not also on the primary DT. This approach can handle conjugate trees as well as non-conjugate trees (link disjoint trees that are rooted at different rBridges).
4. Section 3.2.2: A variation on the section 3.2.1.1b approach, but without Affinity sub-TLV advertisement. Once the primary DT is constructed, costs for links on the primary DT are multiplied by a fixed multiplier to prevent them from being selected in a subsequent SPF run, unless there is no other choice, and the subsequent SPF yields links on the backup DT.

All of the approaches above yield maximally link disjoint trees, when applied as prescribed.

Approach 4 above does not seem to use Affinity sub-TLVs and instead seems to depend upon a network wide agreement on the alternative tree computation algorithm being used.

Approaches 2 and 3 use Affinity sub-TLV on the backup DT, for links that are not already on the primary DT. The primary DT does not appear to use Affinity sub-TLVs. Additionally, from an end-to-end perspective the backup DT comes into picture when the primary DT fails (this is effectively true even in the 1+1 protection mechanism

and in the local protection case), and then again, only until the primary DT is recalculated. Once the primary DT is recalculated, the backup DT is recalculated as well, and can change corresponding to the new primary DT.

draft-ietf-trill-resilient-trees cannot directly prevent/mitigate a parent node shift on the primary DT at a given parent node, and while usage of the Affinity sub-TLV on the backup DT might confer a parent affinity on some nodes on the backup DT, these are not necessarily the nodes on which the network operator may want/prefer an explicit parent affinity. Further, the backup DT is only used on a transient basis, from a forwarding perspective, until the primary DT is recomputed.

However, a parent shift can be triggered by link or node failure. In a situation where both drafts are active in the implementation, failure of a specific link may cause the backup DT to kick in, but when the primary DT is re-calculated, draft-rp-trill-parent-selection-03 can be used to preserve parent-child relationships on the primary DT, to the extent possible, during the re-calculation. So, there does not appear to be a direct functional overlap in the simultaneous usage of these drafts, and it ought to be possible to use both drafts simultaneously, so long as the primary and back-up DTs can be uniquely identified/differentiated.

7. Security Considerations.

The proposal primarily influences tree construction and tries to preserve parent-child relationships in the tree from prior computations of the same tree, without changing any of operational aspects of the protocol. Hence, no new security considerations for TRILL are raised by this proposal.

8. IANA Considerations.

No new registry entries are requested to be assigned by IANA. The Affinity Sub-TLV has been defined in [RFC7176], and this proposal does not change its semantics in any way.

9. Informative 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 (RBrigdes): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <<http://www.rfc-editor.org/info/rfc6325>>.
- [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>>.
- [RFC7783] Senevirathne, T., Pathangi, J., Hudson, J., "Coordinated Multicast Trees (CMT) for Transparent Interconnection of Lots of Links (TRILL)", RFC 7783, February 2016, <<http://datatracker.ietf.org/doc/rfc7783>>
- [RFC4971] Vasseur, JP., Shen, N., Aggarwal, R., "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007, <<http://datatracker.ietf.org/doc/rfc4971>>

Author's Address:

R. Parameswaran,
Brocade Communications, Inc.
120 Holger Way,
San Jose, CA 95134.

Email: parameswaran.r7@gmail.com

Copyright and IPR Provisions

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

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