INTERNET-DRAFT
Updates: 6325, 6361, 7173
Intended status: Proposed Standard
Expires: January 5, 2016

Donald Eastlake
Huawei
Dacheng Zhang
Alibaba

July 6, 2015

TRILL: Link Security
<draft-eastlake-trill-link-security-01.txt>

Abstract

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

[This is a early partial draft.]

Status of This Memo

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

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html. The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Table of Contents

1. Introduction

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

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

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

[TRILL-IP] and other future drafts are expected to cover TRILL security over IP links or other TRILL over X links as specified in the future for technology X.

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

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

[This is a partial early draft.]

1.1 Encryption Requirement and Adjacency

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

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

1.2 Terminology and Acronyms

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

HKDF: Hash based Key Derivation Function [RFC5869].

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

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

MPLS: Multi-Protocol Label Switching.

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

RBridge: An alternative name for a TRILL switch.

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

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

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

2. Link Security Default Keying

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

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

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

3. Link Security Specifics

The following subsection discuss TRILL link security for various technologies.

3.1 Ethernet Links

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

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

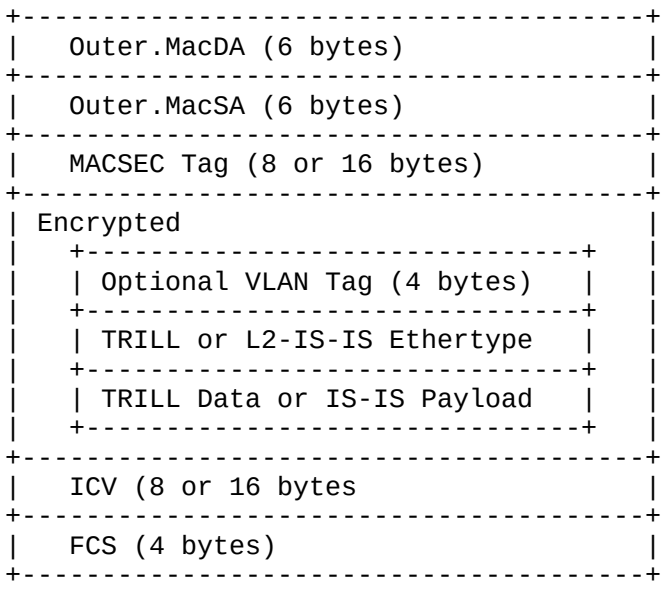
3.1.1 Between TRILL Switches

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

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

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

Of course there may be other Ethernet control frames, such as link aggregation control messages or priority based flow control messages, that would also be sent within MACSEC. Typically only the [802.1X] messages used to establish and maintain MACSEC are sent unsecured.



Figures 1. MACSEC Between TRILL Switch Ports

Outer.MacDA: 48-bit destination MAC address

Outer.MacSA: 48-bit source MAC address

MACSEC Tag: See further description below.

Encrypted: The encrypted data

ICV: The MACSEC Intergrity Check Value

FCS: Frame Check Sequence.

The strucutre of a MACSEC Tag is as follows:

tbd ...

3.1.1.1 Ethernet Link Security Maintenance

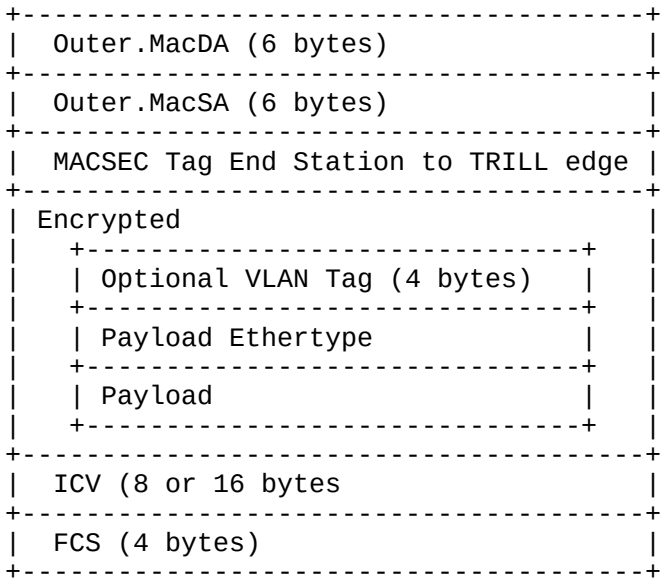
[802.1X] is used to establish keying and algorithms for Ethernet link security ... tbd ...

3.1.2 Ethernet Security to End Stations

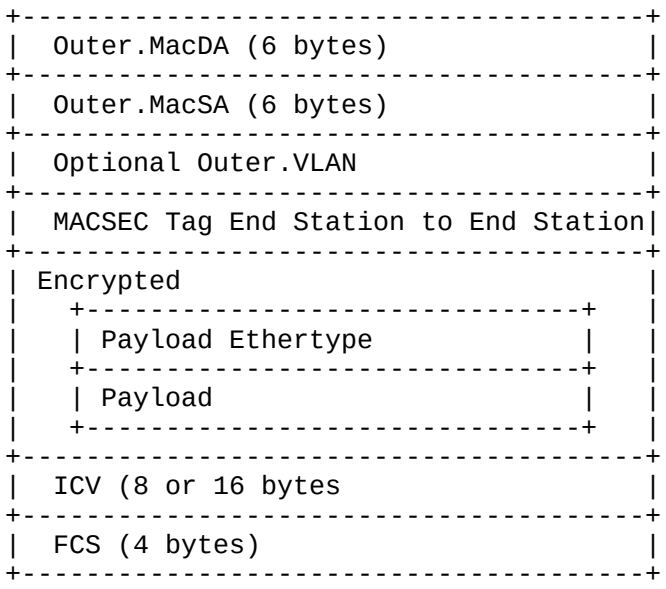
MACSEC may be used between end stations and their adjacent TRILL switch(es) or end-to-end between end stations or both. Since TRILL does not impose administrative requirements on end stations, the choice of keying and crypto suite are beyond the scope of this document.

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

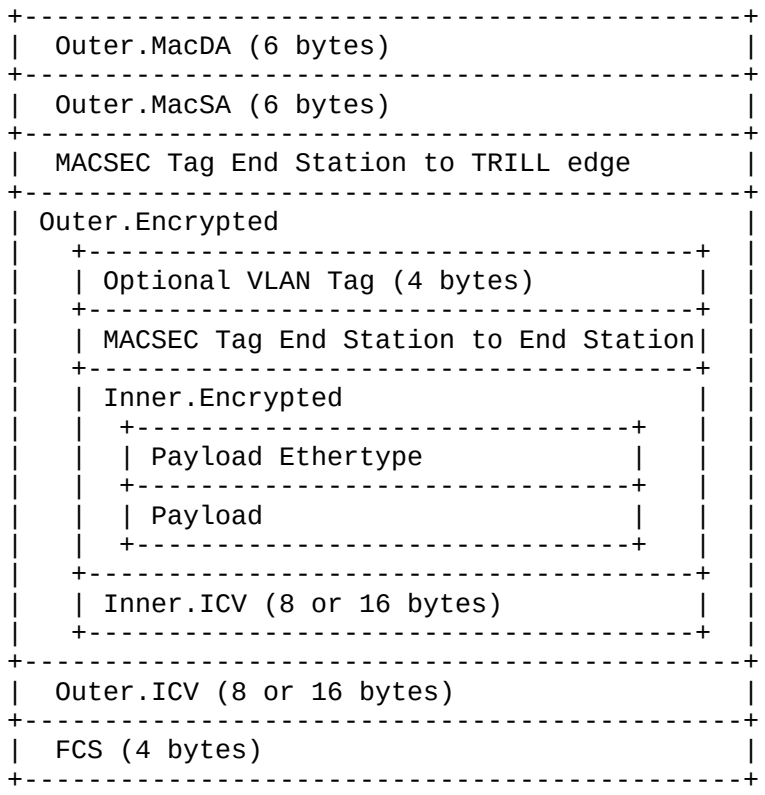
The Figure below show an Ethernet frame between a TRILL switch and the adjacent edge RBridge secured by MACSEC.



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



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



3.2 PPP Links

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

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

3.3 Pseudowire Links

TRILL transport over pseudowires is specified in [RFC7173].

No native security is provided for pseudowires as such; however, they

are, by definition, carried by some PSN (Packet Switched Network). Link security must be provided by this PSN or by lower level protocols. This PSN is typically an MPLS or IP PSN.

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

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

4. Edge-to-Edge Security

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

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

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

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

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

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

more tbd ...

5. Security Considerations

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

For general TRILL Security Considrations, see [RFC6325].

6. IANA Considerations

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

Normative References

[802.1AE] - IEEE Std 802.1AE-2006, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security, 18 August 2006.

[802.1AEbn] - IEEE Std 802.1AEbn-2011, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security / Galois Counter Mode - Advanced Encryption Standard - 256 (GCM-AES-256) Cipher Suite, 14 October 2011.

[802.1AEbw] - IEEE Std 802.1AEbw-2014, IEEE Standard for Local and metropolitan networks / Media Access Control (MAC) Security / Extended Packet Numbering, 12 February 2014

[RFC20] - Cerf, V., "ASCII format for network interchange", STD 80, RFC 20, October 1969, <http://www.rfc-editor.org/info/rfc20>.

[RFC1661] - Simpson, W., Ed., "The Point-to-Point Protocol (PPP)", STD 51, RFC 1661, July 1994, <http://www.rfc-editor.org/info/rfc1661>.

[RFC1968] - Meyer, G., "The PPP Encryption Control Protocol (ECP)", RFC 1968, June 1996, <http://www.rfc-editor.org/info/rfc1968>.

[RFC2119] -Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC5226] - T. Narten and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs," BCP 26 and RFC 5226, May 2008

[RFC5310] - Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.

[RFC5869] - Krawczyk, H. and P. Eronen, "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)", RFC 5869, May 2010, <http://www.rfc-editor.org/info/rfc5869>

[RFC6234] - Eastlake 3rd, D. and T. Hansen, "US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)", RFC 6234, May 2011, <http://www.rfc-editor.org/info/rfc6234>.

[RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011, <http://www.rfc-editor.org/info/rfc6325>.

[RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, August 2011, <http://www.rfc-editor.org/info/rfc6361>.

[RFC7173] - Yong, L., Eastlake 3rd, D., Aldrin, S., and J. Hudson, "Transparent Interconnection of Lots of Links (TRILL) Transport Using Pseudowires", RFC 7173, May 2014, <http://www.rfc-editor.org/info/rfc7173>.

[RFC7177] = Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014, <http://www.rfc-editor.org/info/rfc7177>.

Informative References

[RFC1994] - Simpson, W., "PPP Challenge Handshake Authentication Protocol (CHAP)", RFC 1994, August 1996, <http://www.rfc-editor.org/info/rfc1994>.

[RFC3748] - B. Aboba, et al., "Extensible Authentication Protocol (EAP)," RFC 3748, June 2004

[RFC7258] - Farrell, S. and H. Tschofenig, "Pervasive Monitoring Is an Attack", BCP 188, RFC 7258, May 2014, <http://www.rfc-editor.org/info/rfc7258>.

[RFC7435] - Dukhovni, V., "Opportunistic Security: Some Protection Most of the Time", RFC 7435, December 2014, <http://www.rfc-editor.org/info/rfc7435>.

[rfc7180bis] - Eastlake, D., Zhang, M., Perlman, R. Banerjee, A., Ghanwani, A., and S. Gupta, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, work in progress.

[TRILL-IP] -

Acknowledgments

The authors thank the following for their comments and help:

tbd

Authors' Addresses

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

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

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

Email: dacheng.zdc@alibaba-inc.com

Copyright and IPR Provisions

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

Margaret Cullen
Painless Security
Donald Eastlake
Mingui Zhang
Huawei
Dacheng Zhang
Alibaba

Expires: January 5, 2016
July 6, 2015

Transparent Interconnection of Lots of Links (TRILL) over IP
<draft-ietf-trill-over-ip-03.txt>

Abstract

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

Status of This Document

Table of Contents

Table of Contents (continued)

1. Introduction

TRILL switches (RBridges) are devices that implement the IETF TRILL protocol [RFC6325] [RFC7177] [rfc7180bis].

RBridges provide transparent forwarding of frames within an arbitrary network topology, using least cost paths for unicast traffic. They support not only VLANs and Fine Grained Labels [RFC7172] but also multipathing of unicast and multi-destination traffic. They use IS-IS link state routing and encapsulation with a hop count.

Ports on different RBridges can communicate with each other over various link types, such as Ethernet [RFC6325], pseudowires [RFC7173], or PPP [RFC6361].

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

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

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

To support the use of TRILL over IP encapsulation with good fast path hardware support, a method is provided for agreement between adjacent TRILL switches as to what encapsulation to use. This document updates [RFC7177] and [RFC7178] as described in Section 7 by redefining an interval of RBridge Channel protocol numbers to indicate encapsulation method support for TRILL over IP and by making adjacency between TRILL over IP ports dependent on having a method of encapsulation in common.

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].

3. Use Cases for TRILL over IP

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

3.1 Remote Office Scenario

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

3.2 IP Backbone Scenario

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

3.3 Important Properties of the Scenarios

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

3.3.1 Security Requirements

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

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

3.3.2 Multicast Handling

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

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

3.3.3 RBridge Neighbor Discovery

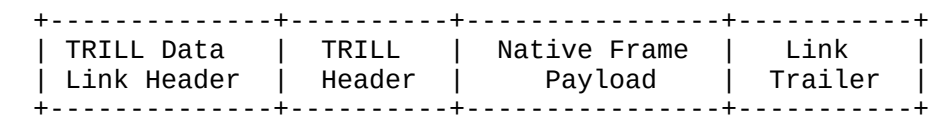
In the IP Backbone Scenario, RBridges that use TRILL over IP can use the normal TRILL IS-IS Hello mechanisms to discover the existence of other RBridges on the link [RFC7177], and to establish authenticated communication with those RBridges.

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 RBridge or a border device at a Remote Office to initiate an IPsec session and subsequent TRILL exchanges with a TRILL over IP-enabled RBridge attached to the TRILL campus.

4. TRILL Packet Formats

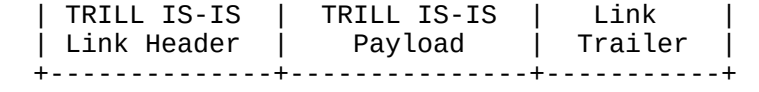
To support the TRILL base protocol standard [RFC6325], two types of packets are transmitted between RBridges: TRILL Data packets and TRILL IS-IS packets.

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



Where 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 TRILL Data Link Header ends with the TRILL Ethertype while the TRILL IS-IS Link Header ends with the L2-IS-IS Ethertype; on the other hand, over PPP, there are no Ethertypes but PPP protocol code points are included that distinguish TRILL Data from TRILL IS-IS.

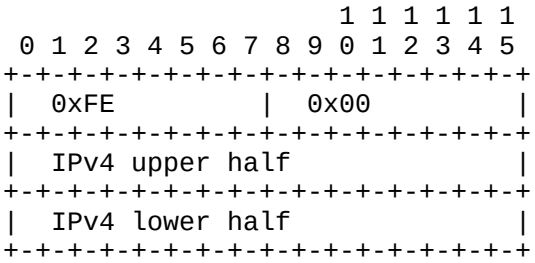
In TRILL over IP, we will use IP (v4 or v6) in the link header. (On the wire, the IP header will be preceeded by the lower layer protocol that is carrying IP, such as Ethernet.) However, there are several IP based encapsulations usable for TRILL over IP as further discussed in Section 7 that differ in exactly what appears after the IP header and before the TRILL header.

5. Some Link Protocol Specifics

TRILL Data packets can be unicast to a specific RBridge or multicast to all RBridges on a link. TRILL IS-IS packets are always multicast to all other RBridge on the link (except for MTU PDUs, which may be unicast [RFC7177]. On Ethernet links, the Ethernet multicast address All-RBridges is used for TRILL Data and All-IS-IS-RBridges for TRILL IS-IS.

To properly handle TRILL base protocol packets on a TRILL over IP link in the general case, either native IP multicast mode must be used on that link, or multicast must be simulated using serial IP unicast, as discussed in Section 8. (Of course, if the IP link happens to actually be point-to-point no special provision is needed for handling multicast addressed packets.)

In TRILL Hello PDUs used on TRILL IP links, the IP addresses of the connected IP ports are their real 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 common 48-bit IS-IS SNPAs, for 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] on that link. This synthetic SNPA derived from an IPv4 address is as follows:



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, at the IP level, where TRILL operates on an IP link, TRILL sees only IP stations, not MAC stations, even if the TRILL over IP Link is being carried over Ethernet, so conflict on the link in TRILL IS-IS between a real MAC address and the synethic SNPA (MAC) address as above would be impossible in any case.

6. 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.

6.1 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 an TRILL over IP port discards output packets that fail the possible recursive ingress test (see Section 10.1) unless configured to disable that test.

6.2 Additional per IP Address Cofiguration

The configuration information specified below is per IP address at a TRILL over IP port.

The mapping from TRILL packet priority to Differentiated Services Code Point (DSCP [RFC2474]) can be configured (see Section 10.5).

Each TRILL over IP port has a list of acceptable encapsulations it will use. By default this list consists of one entry for native encapsulation. (See Section 7.) Additional configuration is possible for specific encapsulations as described in Section 6.2.3.

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 6.2.2) or native IP multicast (Section 6.2.1). Each IP address at a TRILL over IP is configured whether or not to use IPsec (Section 6.2.3).

6.2.1 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 appropriate IP multicast

address (IPv4 or IPv6) specified in Section 13.2. The TRILL over IP port may be configured to use a different IP multicast address for multi-destination packets.

6.2.2 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 ports IP address (IPv4 or IPv6). Multi-destination TRILL packets are serially unicast to the addresses in this list. Such a TRILL over IP port will only be able to form adjacencies [RFC7177] with the RBridges at the addresses in this list as those are the only RBridges to which it will send TRILL Hellos.

If this list of destination IP addresses is empty, there is no way to transmit a multi-destination TRILL over IP packet such as a TRILL Hello. Thus it is impossible to achieve adjacency [RFC7177] or if adjacency had been achieved (perhaps the list was non-empty and has just been configured to be empty), no way to maintain such adjacency. Thus, in the empty list case, TRILL Data multi-destination packets cannot be sent and TRILL Data unicast packets will not start flowing or, if they are already flowing, will soon cease, effectively disabling the port.

6.2.3 Encapsulation Specific Configuration

Specific TRILL over IP encapsulation methods may provide for futher configuration as specified in the subsections below.

6.2.3.1 VXLAN Configuration

A TRILL over IP port using VXLAN encapsulation can be configured with a non-default VXLAN Network Identifier (VNI) which is used in that field of the VXLAN header for all TRILL packets sent using the encapsulation and required in all TRILL packets received using the encapsulation. In this case, a TRILL packet received with the wrong VNI is discarded.

A TRILL over IP port using VXLAN encapsulation can also be configured to place the Inner.VLAN or Inner.FGL of a TRILL Data packet being transported in the VNI field.

6.2.3.2 Other Encapsulation Configuration

[Specific configuration for other encapsulation methods will be added here.]

6.2.4 Security Configuration

tbd ...

7. TRILL over IP Encapsulation Formats

There are a variety of TRILL over IP formats possible. In all cases, there must be a method specified, with each format, to distinguish TRILL Data and TRILL IS-IS packets, or that format is not useful for TRILL. The following criteria can be helpful is choosing between different encapsulations:

a) Fast path support - For most applications, it is highly desireable to be able to encapsulate/decpasulate TRILL over IP at line speed so a format where existing or anticipated fast path hardware can do that is best.

b) Ease of multi-pathing - The IP path between TRILL over IP port may include internal equal cost multipath routes so a method of encapsulation that provides variable fields available for existing or anticipated fast path hardware multi-pathing is better.

c) Fragmentaton and robust ID support - tbd

d) Checksum strength - Depending on the particular circumstances of the TRILL over IP link, a checksum provided by the encapsulation may be an important factor. Use of IPsec as provided herein can also provide a strong integrity check.

TRILL over IP adopts a hybrid encapsulation approach by default.

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 with low bandwidth control messages to detetmine 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 encpasulation used for all TRILL Data packets and all other TRILL IS-IS PDUs with the possible exception of IS-IS MTU-probe and MTU-ack PDUs as discussed in Section 7.

Alternatively, the network operator can pre-configure a TRILL over IP port to use a particular encapsulation choosen for their particular network needs and TRILL over IP port capabiliies for all TRILL data and IS-IS packets.

Section 7.1 discusses encapsulation agreement. Section 7.2 discusses TRILL over IP IPsec ESP format, which is independent of encapsulation. Section 7.3 discusses broadcast link encapsulation considerations. The subsequent subsections discuss particular encapsulations.

7.1 Encapsulation Agreement

TRILL Hellos sent out a TRILL over IP port indicate the encapsulations that port is willing to use through the mechanism described in [RFC7178] and [RFC7176]. RBridge Channel Protocol numbers 0xFC0 through 0xFF7 are hereby redefined to be link technology dependent flags that, for TRILL over IP, indicate support for different encapsulations, allowing for up to 24 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 13.3.) "Support" indicates willingness to use that encapsulation for TRILL data and TRILL IS-IS other than Hellos. Even if support is not indicated for native encapsulation, by default support for native encapsulation of TRILL Hellos is assumed.

If no encapsulation support is indicated in a TRILL Hello, then the port from which it was sent is assumed to support only native encapsualtion (see Section 7.4).

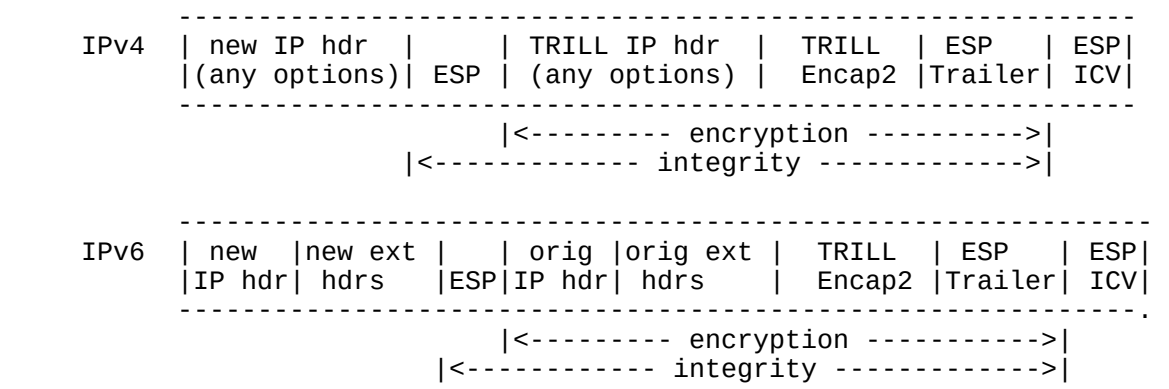
An adjacency is formed between two TRILL over IP ports ONLY if the intersection of the sets of encapsulation methods they support is not null. If that intersection is null, then no adjaceny is formed. In particular, for a TRILL over IP link, the adjacency state machine MUST NOT advance to the Report state unless the ports share an encapsulation [RFC7177].

If any TRILL over IP packet, other than a IS-IS Hello or MTU PDU in native encapsulation, is received in an encapsulation for which support is not being indicated, it MUST be discarded (see Section 7.3).

It expected to normally be the case in a well configured network that all the TRILL over IP ports connected to an IP network that are intended to communicate with each other will support the same encapsulation. But the network will operate correctly if this is not true.

7.2 IPsec ESP Format

TRILL over IP link security uses IPsec Encapsulating Security Protocol (ESP) in tunnel mode [RFC4303]. Since TRILL over IP always starts with an IP Header (on the wire this appears right after any required Layer 2 header), the modifications when IPsec is in effect are independent of the TRILL over IP encapsulation fields that occur after that IP Header and before the TRILL Header. The resulting packet formats are as follows for IPv4 and IPv6:



The "TRILL Encap2" above includes whatever additional fields are required by the encapsulation in use followed by the TRILL Header and then the native frame payload (see Section 4).

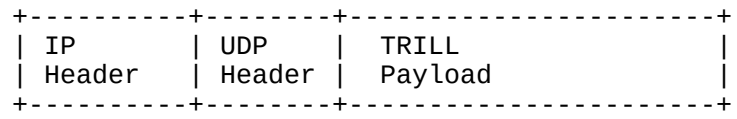
This architecture permits the ESP tunnel termination to be separated from the TRILL over IP RBridge port and, for example, placed at a physical or administrative security boundary.

7.3 Broadcast Link Encapsulation Considerations

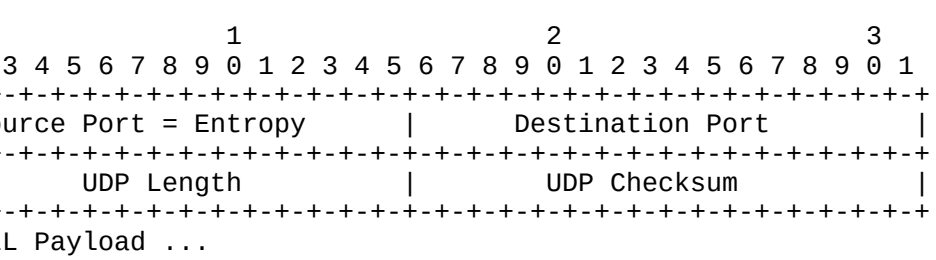
It is possible for the Hellos from a TRILL over IP port P1 to establish adjacency with multiple other TRILL over IP ports (P2, P3, ...) forming a broadcast link. In a well configured network one would expect such multiple other IP ports to support the same encapsulation but, if P1 supports multiple encapsulations, it is possible that P2 and P3, for example, do not have an encapsulation in common that is supported by P1. IS-IS can handle such non-transitive adjacencies which are reported as specified in [RFC7177]. If serial IP unicast is being used by P1, it can use different encapsulatons for different transmission. If native IP multicast is being used by P1, it will have to send one transmission per encapsulation method by which it has an adjanceny on the link. (It is for this reason that a TRILL over IP port MUST discard any packet received with the wrong encapsulation. Otherwise, packets would be duplicated.)

7.4 Native Encapsulaton

The mandatory to implement "native encapsulaton" format of a TRILL over IP packet, when used without security, is TRILL over UDP as shown below.



Where the UDP Header is as follows:



Source Port - see Section 10.2

Destination Port - indicates TRILL Data or IS-IS, see Section 13

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 Routeing Protocol Discriminator byte (thus not including the L2-IS-IS Ethertype) for TRILL IS-IS packets.

7.5 VXLAN Encapsulation

VXLAN [RFC7348] IP encapsulation of TRILL looks, on the wire, as 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 7.2). The VXLAN header after the outer UDP header adds a 24 bit Virtual Network Identifier. The Ethernet header after the VXLAN header and before the TRILL header provides an Ethertype field that distinguishes TRILL data from TRILL IS-IS; however, the destination and source MAC addresses in this inner Ethernet header before the TRILL header are not used and represent 12 wasted bytes.

A TRILL over IP port using VXLAN encapsulation by default uses a VNI of 1 but can be configured as described in Section 6.2.3.1.

7.6 Other Encpaulsations

[Additional encpasulations will be added here as additional subsections.]

8. 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 13.2); however, a TRILL over IP port may be configured (see Section 6) 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 that case the outer IP header shows the IP unicast even through the TRILL header has the M bit set to one to indicate multi-destination. Serial unicast configuration is necessary if the TRILL over IP port is connected to an IP network that does not support IP multicast. In any case, unicast TRILL packets are sent by unicast IP.

When a TRILL over IP port is using IP multicast, it MUST periodically transmit appropriate IGMP (IPv4 [RFC3376] or MLD (IPv6 [RFC2710]) packets so that the TRILL multicast IP traffic will be sent to it.

Although TRILL fully supports broadcast links with more than 2 RBridges connected to the link, even where native IP multicast is available, there may be good reasons for configuring TRILL over IP ports to use serial unicast. In some networks, unicast is more reliable than multicast. If multiple unicast connections between parts of a TRILL campus are configured, TRILL will in any case spread traffic across them, treating them as parallel links, and appropriately fail over traffic if a link ceases to operate or incorporate a new link that comes up.

9. Use of IPsec

All RBridges that support TRILL over IP MUST implement IPsec [RFC4301] and support the use of IPsec Encapsulating Security Protocol (ESP [RFC4303]) 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, implementors may elect use IS-IS security to protect TRILL IS-IS packets. However, in this case, the IPsec session still MUST be fully established before any data packets transmission since IS-IS security does not provide any protection to data packets.

9.1 Default Keys

The default pre-shared keys for IPsec usage are derived as follows:

HMAC-SHA256 ("TRILL IP"| IS-IS-shared key)

In the above "|" indicates concatenation, HMAC-SHA256 is as described in [FIPS180] [RFC6234] and "TRILL IP" is the eight byte US ASCII [RFC0020] string indicated. "IS-IS-shared key" is a link (or wider scope) IS-IS key usable for IS-IS security of link local IS-IS local PDUs such as Hello, CSNP, and PSNP. With [RFC5310] there could be multiple keys identified with 16-bit key IDs. In this case, the Key ID of IS-IS-shared key is also used to identify the derived key.

Although we are using pre-shared keys at the IPsec level, the IS-IS-shared keys from which they are derived expire and can be updated as described in RFC 5310. The derived keys MUST expire within the lifetime as the IS-IS-shared keys from which they were derived.

9.2 Mandatory-to-Implement Algorithms

All RBridges that support TRILL over IP MUST implement the following algorithms for IPsec ESP, as recommended in [RFC4308]:

Protocol	ESP [RFC4303]
ESP encryption	AES with 128-bit keys in CBC mode [RFC3602]
ESP integrity	AES-XCBC-MAC-96 [RFC3566]

10. Transport Considerations

This section discusses a variety of transport considerations.

10.1 Recursive Ingress

TRILL is designed to transport end station traffic to and from end stations over IEEE 802.3 and IP is frequently transported over IEEE 802.3 or similar protocols. Thus, an end station native data frame EF might get TRILL ingressed to TRILL(EF) which was then sent oout a TRILL over IP over 802.3 port resulting in an 802.3 frame of the form 802.3(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, etc. The packet might get discarded if it got too large but if fragmentation is enabled, it would just keep getting split into fragments that would continue to loop and grow and re-fragment until the path was saturated with junk and packets were being discarded due to queue overflow. The TRILL Header TTL would provide no protection because each TRILL ingress adds a new Header and TTL.

To protect against this scenario, a TRILL over IP port MUST by, default, test whether a TRILL packet it is about to transmit is, in fact a TRILL ingress of a TRILL over IP over 802.3 or the like packets. That is, is it of the form TRILL(802.3(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.

10.2 Fat Flows

For the purpose of load balancing, it is worthwhile to consider how to transport the TRILL packets over the Equal Cost Multiple Paths (ECMPs) existing internal to the IP path between two TRILL over IP ports.

The ECMP election for the IP traffic could be based, at least for IPv4, on the quintuple of the outer IP header { Source IP, Destination IP, Source Port, Destination Port, and IP protocol }. Such tuples, however, could be exactly the same for all TRILL Data packets between two RBridge ports, even if there is a huge amount of data being sent between a variety of ingress and egress RBridges. Therefore, in order to better support ECMP, a RBridge SHOULD set the Source Port as an entropy field for ECMP decisions. (This idea is also introduced in [gre-in-udp].) For example, for TRILL Data this

entropy field could be based on the Inner.MacDA, Inner.MacSA, and Inner.VLAN or Inner.FGL.

10.3 Congestion Considerations

Section 3.1.3 of [RFC5405] discussed the congestion implications of UDP tunnels. As discussed in [RFC5405], because other flows can share the path with one or more UDP tunnels, congestion control [RFC2914] needs to be considered.

The default initial determination of the TRILL over IP encapsulation to be used through the exchange of TRILL IS-IS Hellos is a low bandwidth process. Hellos are not permitted to be sent any more often than once per second, and so are unlikely to cause congestion.

One motivation for including UDP in a TRILL encapsulation is to improve the use of multipath (such as ECMP) in cases where traffic is to traverse routers which are able to hash on UDP Port and IP address. In many cases this may reduce the occurrence of congestion and improve usage of available network capacity. However, it is also necessary to ensure that the network, including applications that use the network, responds appropriately in more difficult cases, such as when link or equipment failures have reduced the available capacity.

The impact of congestion must be considered both in terms of the effect on the rest of the network of a UDP tunnel that is consuming excessive capacity, and in terms of the effect on the flows using the UDP tunnels. The potential impact of congestion from a UDP tunnel depends upon what sort of traffic is carried over the tunnel, as well as the path of the tunnel.

TRILL is used to carry a wide range of traffic. In many cases TRILL is used to carry IP traffic. IP traffic is generally assumed to be congestion controlled, and thus a tunnel carrying general IP traffic (as might be expected to be carried across the Internet) generally does not need additional congestion control mechanisms. As specified in [RFC5405]:

"IP-based traffic is generally assumed to be congestion- controlled, i.e., it is assumed that the transport protocols generating IP-based traffic at the sender already employ mechanisms that are sufficient to address congestion on the path. Consequently, a tunnel carrying IP-based traffic should already interact appropriately with other traffic sharing the path, and specific congestion control mechanisms for the tunnel are not necessary".

For this reason, where TRILL is sent using UDP and used to carry IP traffic that is known to be congestion controlled, the UDP paths MAY

be used across any combination of a single or cooperating service providers or across the general Internet.

However, TRILL is also used to carry traffic that is not necessarily congestion controlled. For example, TRILL may be used to carry traffic where specific bandwidth guarantees are provided.

In such cases congestion may be avoided by careful provisioning of the network and/or by rate limiting of user data traffic. Where TRILL is carried, directly or indirectly, over UDP over IP, the identity of each individual TRILL flow is in general lost.

For this reason, where the TRILL traffic is not congestion controlled, TRILL over UDP/IP MUST only be used within a single service provider that utilizes careful provisioning (e.g., rate limiting at the entries of the network while over-provisioning network capacity) to ensure against congestion, or within a limited number of service providers who closely cooperate in order to jointly provide this same careful provisioning. As such, TRILL over UDP/IP MUST NOT be used over the general Internet, or over non-cooperating service providers, to carry traffic that is not congestion-controlled.

Measures SHOULD be taken to prevent non-congestion-controlled TRILL over UDP/IP traffic from "escaping" to the general Internet, for example the following:

a. Physical or logical isolation of the TRILL over IP links from the general Internet.

b. Deployment of packet filters that block the UDP ports assigned for TRILL-over-UDP.

c. Imposition of restrictions on TRILL over UDP/IP traffic by software tools used to set up TRILL over UDP paths between specific end systems (as might be used within a single data center).

d. Use of a "Managed Circuit Breaker" for the TRILL traffic as described in [circuit-breaker].

10.4 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 RBridges in the campus) through the originatingLSPBufferSize TLV [RFC6325] [RFC7177]. The campus minimum MTU, denoted Sz, is then

established by taking the minimum of this advertised MTU for all RBridges in the campus. Links that do not meet the Sz MTU are not included in the routing topology. This protects the operation of IS-IS from links that would be unable to accommodate some LSPs.

A method of determining originatingLSPBufferSize for an RBridge with one or more TRILL over IP ports is described in [rfc7180bis]. However, if an IP link either can accommodate jumbo frames or is a link on which IP fragmentation is enabled and acceptable, then it is unlikely that the IP link will be a constraint on the originatingLSPBufferSize of an 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, there may be links that meet the minimum MTU for the IP protocol (1,280 bytes for IPv6, theoretically 68 bytes for IPv4) on which it would be necessary to enable fragmentation for TRILL use.

The optional use of TRILL IS-IS MTU PDUs, as specified in [RFC6325] and [RFC7177] can provide added assurance of the actual MTU of a link.

10.5 QoS Considerations

Within TRILL, priority is indicated by a three bit (0 through 7) priority field in TRILL data packets and by configuration for TRILL IS-IS packets. When TRILL packets are sent on a TRILL over IP link, this priority is mapped to a Differential Services Code Point (DSCP [RFC2474] Section 4.2.2).

The default mapping, which may be configured per TRILL over IP port, is an follows. Note that, to provide a potentially lower priority service than the default 0, priority 1 is considered lower priority than 0. So the priority squence from lower to higher priority is 1, 0, 2, 3, 4, 5, 6, 7.

TRILL Priority	DiffServ Field (Binary/decimal)
0	00100000 / 8
1	00000000 / 0
2	01000000 / 16
3	01100000 / 24
4	10000000 / 32
5	10100000 / 40
6	11000000 / 48
7	11100000 / 56

11. Middlebox Considerations

TBD ...

12. Security Considerations

TRILL over IP is subject to all of the security considerations for the base TRILL protocol [RFC6325]. In addition, there are specific security requirements for different TRILL deployment scenarios, as discussed in the "Use Cases for TRILL over IP" section above.

This document specifies that all RBridges that support TRILL over IP MUST implement IPsec, and makes it clear that it is both wise and good to use IPsec in all cases where a TRILL over IP link will traverse a network that is not under the same administrative control as the rest of the TRILL campus or is not physically secure. IPsec is necessary, in these cases to protect the privacy and integrity of data traffic.

TRILL over IP is compatible with the use of IS-IS Security [RFC5310], which can be used to authenticate RBridges before allowing them to join a TRILL campus. This is sufficient to protect against rogue RBridges, but is not sufficient to protect data packets that may be sent in IP outside of the local network, or even across the public Internet. To protect the privacy and integrity of that traffic, use IPsec.

In cases were 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. In cases where both types of security are enabled, by default, a key derived from the IS-IS key will be used for IPsec.

13. IANA Considerations

IANA considerations are given below.

13.1 Port Assignments

IANA has allocated the following destination UDP Ports for the TRILL IS-IS and Data channels:

UDP Port	Protocol
(TBD1)	TRILL IS-IS Channel
(TBD2)	TRILL Data Channel

13.2 Multicast Address Assignments

IANA has allocated one IPv4 and one IPv6 multicast address, as shown below, which correspond to the All-RBridges and All-IS-IS-RBridges multicast MAC addresses that the IEEE Registration Authority has assigned for TRILL. Because the low level hardware MAC address dispatch considerations for TRILL over Ethernet do not apply to TRILL over IP, one IP multicast address for each version of IP is sufficient.

(Values recommended to IANA in square brackets)

Name	IPv4	IPv6
All-RBridges	TBD3[233.252.14.0]	TBD4[FF0X:0:0:0:0:0:0:205]

Note: when these IPv4 and IPv6 multicast addresses are used and the resulting IP frame is sent over Ethernet, the usual IP derived MAC address is used.

[Need to discuss scopes for IPv6 multicast (the "X" in the addresses) somewhere. Default to "site" scope but MUST be configurable?]

13.3 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 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	
0x100-0xFBF	RFC Required	allocation of a single value
0x100-0xFBF	IESG Approval	allocation of multiple values
0xFC0 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
  0x004-0xFBF Unassigned
  0xFC0-0xFF7 (link technology dependent, see subregistry)

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

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

Flag	Meaning
0xFC0	Native encapsulation supported
0xFC1	VXLAN encapsulation supported
0xFC2-0xFF7	Unassigned

Normative References

[IS-IS] - "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, 2002".

[RFC0020] - Cerf, V., "ASCII format for network interchange", STD 80, RFC 20, DOI 10.17487/RFC0020, October 1969, <http://www.rfc-editor.org/info/rfc20>.

[RFC0768] - Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <http://www.rfc-editor.org/info/rfc768>.

[RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC2474] - Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <http://www.rfc-editor.org/info/rfc2474>.

[RFC2710] - Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, DOI 10.17487/RFC2710, October 1999, <http://www.rfc-editor.org/info/rfc2710>.

[RFC2914] - Floyd, S., "Congestion Control Principles", BCP 41, RFC 2914, DOI 10.17487/RFC2914, September 2000, <http://www.rfc-editor.org/info/rfc2914>.

[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>.

[RFC3566] - Frankel, S. and H. Herbert, "The AES-XCBC-MAC-96 Algorithm and Its Use With IPsec", RFC 3566, DOI 10.17487/RFC3566, September 2003, <http://www.rfc-editor.org/info/rfc3566>.

[RFC3602] - Frankel, S., Glenn, R., and S. Kelly, "The AES-CBC Cipher Algorithm and Its Use with IPsec", RFC 3602, DOI 10.17487/RFC3602, September 2003, <http://www.rfc-editor.org/info/rfc3602>.

[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>.

[RFC4308] - Hoffman, P., "Cryptographic Suites for IPsec", RFC 4308, DOI 10.17487/RFC4308, December 2005, <http://www.rfc-editor.org/info/rfc4308>.

[RFC5405] - Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <http://www.rfc-editor.org/info/rfc5304>.

[RFC5310] - Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <http://www.rfc-editor.org/info/rfc5310>.

[RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <http://www.rfc-editor.org/info/rfc6325>.

[RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, DOI 10.17487/RFC7176, May 2014, <http://www.rfc-editor.org/info/rfc7176>.

[RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, DOI 10.17487/RFC7177, May 2014, <http://www.rfc-editor.org/info/rfc7177>.

[RFC7178] - Eastlake 3rd, D., Manral, V., Li, Y., Aldrin, S., and D. Ward, "Transparent Interconnection of Lots of Links (TRILL): RBridge Channel Support", RFC 7178, DOI 10.17487/RFC7178, May 2014, <http://www.rfc-editor.org/info/rfc7178>.

[RFC7348] - Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <http://www.rfc-editor.org/info/rfc7348>.

[rfc7180bis] - Eastlake, D., et al, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, work in progress.

[FIPS180] - "Secure Hash Standard (SHS)", United States of American, National Institute of Science and Technology, Federal Information Processing Standard (FIPS) 180-4, March 2012.

Informative References

[RFC6234] - Eastlake 3rd, D. and T. Hansen, "US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)", RFC 6234, DOI 10.17487/RFC6234, May 2011, <http://www.rfc-editor.org/info/rfc6234>.

[RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, DOI 10.17487/RFC6361, August 2011, <http://www.rfc-editor.org/info/rfc6361>.

[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>.

[gre-in-udp] - Crabbe, E., Yong, L., and X. Xu, "Generic UDP Encapsulation for IP Tunneling", draft-yong-tsvwg-gre-in-udp-encap, work in progress.

[circuit-breaker] - Fairhurst, G., "Network Transport Circuit Breakers", draft-ietf-tsvwg-circuit-breaker, work in progress.

Acknowledgements

The following people have provided useful feedback on the contents of this document: Sam Hartman, Adrian Farrel.

Some material in Section 10.2 is derived from draft-ietf-mpls-in-udp by Xiaohu Xu, Nischal Sheth, Lucy Yong, Carlos Pignataro, and Yongbing Fan.

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

Authors' Addresses

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

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

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

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

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

EMail: zhangmingui@huawei.com

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

Email: dacheng.zdc@alibaba-inc.com

Copyright, Disclaimer, and Additional IPR Provisions

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

Mohammed Umair
Kingston Smiler
IP Infusion
Donald Eastlake 3rd
Lucy Yong
Huawei Technologies

Expires: January 7, 2016
July 6, 2015

TRILL Transparent Transport over MPLS
draft-muks-trill-transport-over-mpls-00

Abstract

This document specifies how to interconnect Transparent Interconnection of Lots of links (TRILL) sites belonging to a tenant that are separated geographically over an MPLS domain. This draft addresses two problems 1) Providing connection between more than two TRILL sites that are separated by an MPLS provider network using [RFC7173] 2) Providing connection between TRILL sites belonging to a tenant over a MPLS provider network

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Copyright and License Notice

Table of Contents

1 Introduction

The IETF Transparent Interconnection of Lots of Links (TRILL) protocol [RFC6325] [RFC7177] [RFC7180bis] provides transparent forwarding in multi-hop networks with arbitrary topology and link technologies using a header with a hop count and link-state routing. TRILL provides optimal pair-wise forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. Intermediate Systems (ISs) implementing TRILL are called Routing Bridges(RBridges) or TRILL Switches

This draft, in conjunction with [RFC7173], address two problems

1) Providing connection between more than two TRILL sites of a single TRILL network that are separated by an MPLS provider network using [RFC7173]. (Herein also called as problem statement 1.)

2) Providing connection between TRILL sites belongs to a tenant/tenants over a MPLS provider network. (Herein also called as problem statement 2.)

A tenant is the administrative entity on whose behalf one or more customers and their associated services are managed. Here Customer refers to TRILL campus not Data Label.

A key multi-tenancy requirement is traffic isolation so that one tenant’s traffic is not visible to any other tenant. This draft also addresses the problem of multi-tenancy by isolating one tenant’s traffic from the other.

1.1 Terminology

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

Acronyms used in this document include the following:

AC - Attachment Circuit [RFC4664]

Data Label - VLAN or FGL

ECMP - Equal Cost Multi Path

FGL - Fine-Grained Labeling [RFC7172]

IS-IS	- Intermediate System to Intermediate System [IS-IS]
LDP	- Label Distribution Protocol
LAN	- Local Area Network
MPLS	- Multi-Protocol Label Switching
PE	- Provider Edge Device
PPP	- Point-to-Point Protocol [RFC1661]
PSN	- Packet Switched Network
PW	- Pseudowire [RFC4664]
TIR	- TRILL Intermediate Router [Devices where Pseudowire starts and Terminates]
TRILL	- Transparent Interconnection of Lots of Links OR Tunneled Routing in the Link Layer
TRILL Site	- A part of a TRILL campus that contains at least one RBridge.
VLAN	- Virtual Local Area Network
VPLS	- Virtual Private LAN Service
VPTS	- Virtual Private TRILL Service
VSI	- Virtual Service Instance [RFC4664]
VTSD	- Virtual TRILL Switch Domain A Virtual RBridge which segregates one tenant’s TRILL database as well as traffic from the other.
WAN	- Wide Area Network

2. TRILL Over MPLS Model

TRILL Over MPLS can be achieved by two different ways.

a) VPLS Model for TRILL
b) VPTS Model/TIR Model

Both these models can be used to solve the problem statement 1 and 2. Herein the VPLS Model for TRILL is also called Model 1 and the VPTS Model/TIR Model is also called Model 2.

3. VPLS Model

Figure 1 shows the topological model of TRILL over MPLS using VPLS model. The PE routers in the below topology model should support all the functional Components mentioned in [RFC4664].

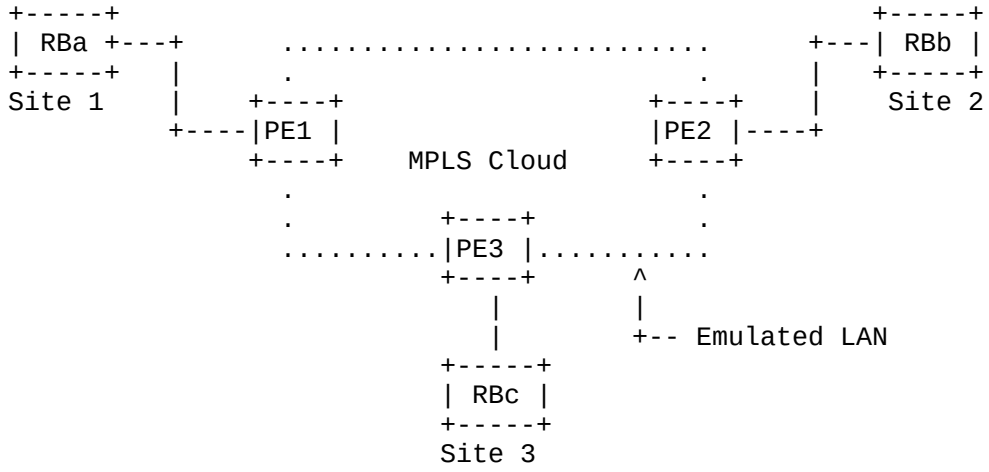
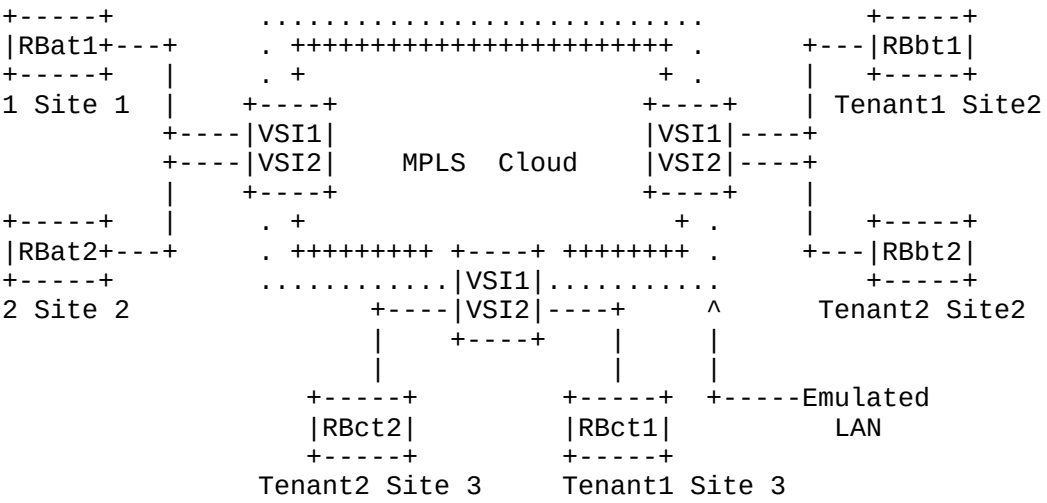


Figure 1: Topological Model of TRILL over MPLS connecting three TRILL Sites

Figure 2 below shows the topological model of TRILL over MPLS to connect multiple TRILL sites belonging to a tenant (tenant here is a campus, not a Data label). VSI1 and VSI2 are two Virtual Service Instances which segregates Tenant1’s traffic from Tenant2’s. VSI1 will maintain its own database for Tenant1, similarly VSI2 will maintain its own database for Tenant2.



.... VSI1 Path
++++ VSI2 Path

Figure 2: Topological Model for VPLS Model
connecting 2 Tenants with 3 sites each

In this model TRILL sites are connected using VPLS-capable PE devices that provide a logical interconnect, such that TRILL RBridges belonging to a specific tenant connected via an single bridged Ethernet. These devices are same as PE devices specified in [RFC4026]. The Attachment Circuit ports of PE Routers are layer 2 switch ports that are connected to the RBridges in a TRILL site. Here each VPLS instance looks like an emulated LAN. This model is similar to connecting different RBridges (TRILL sites) by a layer 2 bridge domain (multi access links) as specified in [RFC6325]. This model doesn’t requires any changes in PE routers to carry TRILL frames, as TRILL frame will be transferred transparently.

3.1. Entities in the VPLS Model

The PE (VPLS-PE) and CE devices are defined in [RFC4026].

The Generic L2VPN Transport Functional Components like Attachment Circuits, Pseudowires, VSI etc. are defined in [RFC4664].

The RB (RBridge) and TRILL Sites are defined in [RFC6325]

3.3. TRILL Adjacency for VPLS model

As specified in section 3 of this document, the MPLS cloud looks like an emulated LAN (also called multi-access link or broadcast link). This results in RBridge of different sites looking like that they are connected to a multi-access link. With such interconnection, the TRILL adjacency over the link is automatically discovered and established through TRILL IS-IS control messages [RFC7177] which is transparently forwarded by the VPLS domain, after doing MPLS encapsulation specified in the section 3.4.

3.4. MPLS encapsulation for VPLS model

MPLS encapsulation over Ethernet pseudowire is specified in [RFC7173] Appendix A, and requires no changes in the frame format.

3.5 Loop Free provider PSN/MPLS.

No explicit handling is required to avoid loop free topology as, Split Horizon technique mentioned in [RFC4664] in the provider PSN network takes care of loop-free topology in the PSN.

3.6. Frame processing.

The PE device transparently process the TRILL control and data frames and procedure to forward the frames are defined in [RFC4664]

4. VPTS Model

The [Virtual Private TRILL Service] VPTS is an L2 TRILL service that emulates TRILL service across a Wide Area Network (WAN). VPTS is similar to what VPLS does for bridge domain. VPLS provides virtual private LAN service for different customers. VPTS provide Virtual Private TRILL service (VPTS) for different TRILL tenants.

Figure 3 shows the topological model of TRILL over MPLS using VPTS. In this model the PE routers are replaced with TIR [TRILL Intermediate Router] and VSI is replaced with VTSD [Virtual TRILL Switch Domain]. The TIR [TRILL Intermediate Router] devices are interconnected via PWs appear as a single emulated TRILL Site with each VTSD inside a TIR equivalent to a RBridge. The TIR devices must be capable of supporting both MPLS and TRILL.

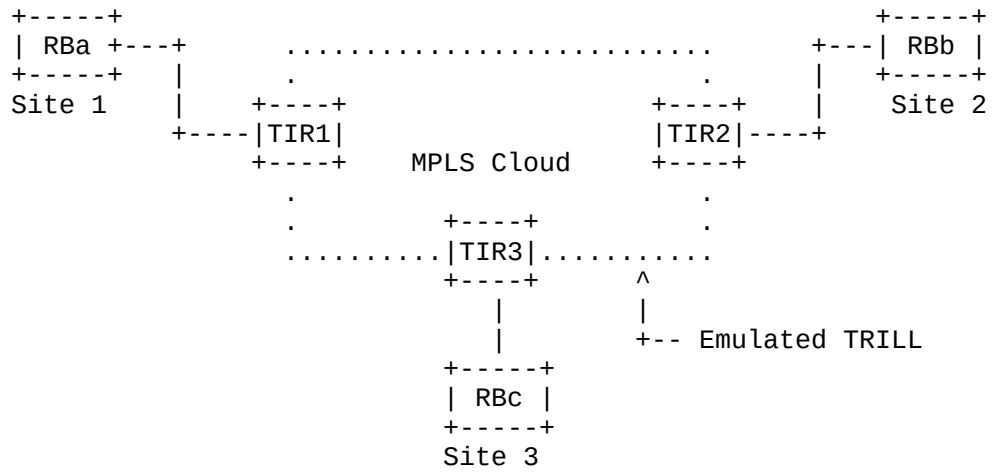


Figure 3: Topological Model of VPTS/TIR
connecting three TRILL Sites

In the above figure (Figure 3) Site1, Site2 and Site3 (running TRILL protocol) are connected to TIR Devices. These TIR devices along with the MPLS cloud looks like an emulated TRILL Site with all the TRILL Sites connecting to MPLS cloud forming a single TRILL campus. Only the PE devices in the MPLS network should be replaced with TIRs that make the intermediate P routers agnostic to the TRILL protocol.

Figure 4 below shows the topological model of TRILL over MPLS to connect multiple TRILL sites belonging to a tenant (tenant here is a campus, not a Data label) using VPTS model. VTSD1 and VTSD2 are two Virtual TRILL Switch Domains (Virtual RBridges) that segregates Tenant1’s traffic from Tenant2’s. VTSD1 will maintain its own TRILL database for Tenant1, similarly VTSD2 will maintain its own TRILL database for Tenant2.

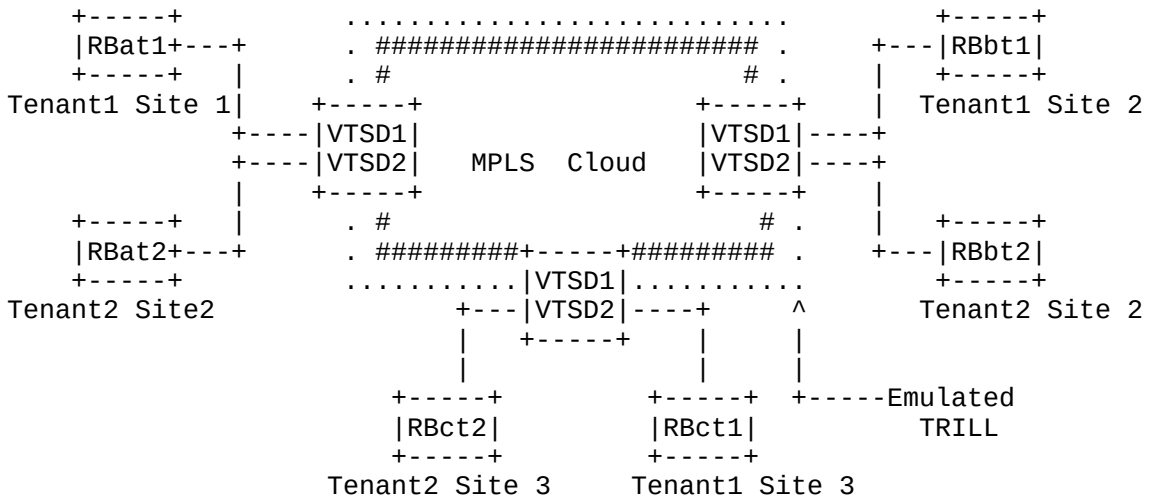


Figure 4: Topological Model of VPTS/TIR
connecting 2 tenants with three TRILL Sites

4.1. Entities in the VPTS Model

The CE devices are defined in [RFC4026].

The Generic L2VPN Transport Functional Components like Attachment Circuits, Pseudowires etc. are defined in [RFC4664].

The RB (RBridge) and TRILL Campus are defined in [RFC6325]

This model introduces two new entities called TIR and VTSD.

4.1.1 TRILL Intermediate Routers [TIR]

The TIRs [TRILL Intermediate Routers] must be capable of running both VPLS and TRILL protocols. TIR devices are superset of VPLS-PE devices which is defined in [RFC4026]. The VSI instance that provides transparent bridging functionality in the PE device is replaced with VTSD in TIR.

4.1.2 Virtual TRILL Switch Domain (VTSD)

The VTSD [Virtual Trill Switch Domain] is similar to VSI (layer 2 bridge) in VPLS model, but this acts as TRILL RBridge. The VTSD is a superset of VSI and must support all the functionality provided by the VSI as defined in [RFC4026]. Along with VSI functionality, the VTSD must be capable of supporting TRILL protocols and form TRILL adjacency. The VTSD must be capable of performing all the operations that a standard TRILL Switch can do.

One VTSD instance per tenant must be maintained, when multiple tenants are connected to the TIR. The VTSD must maintain all the information maintained by the RBridge on a per tenant basis. The VTSD must also take care of segregating one tenant traffic from other.

4.2. TRILL Adjacency for VPLS model

The VTSD must be capable of forming TRILL adjacency with other VTSDs present in its peer VPTS neighbor, and also the RBridges present in the TRILL sites. The procedure to form TRILL Adjacency is specified in [RFC7173] and [RFC7177].

4.3. MPLS encapsulation for VPLS model

MPLS encapsulation over pseudowire is specified in [RFC7173], and requires no changes in the frame format.

4.4 Loop Free provider PSN/MPLS.

This model isn’t required to employ Split Horizon mechanism in the provider PSN network, as TRILL takes care of Loop free topology using Distribution Trees. Any multi-destination frame will traverse a distribution tree path. All distribution trees are calculated based on TRILL base protocol standard [RFC6325] as updated by [RFC7180bis].

4.5. Frame processing.

This section specifies multi-destination and unicast frame processing in VPTS/TIR model.

4.5.1 Multi-Destination Frame processing

Any unknown unicast, multicast or broadcast frames inside VTSD should be

processed or forwarded through any one of the distribution tree's path. If any multi-destination frame is received from the wrong pseudowire at a VTSD, the TRILL protocol running in VTSD should perform a RPF check as specified in [RFC7180bis] and drops the packet.

Pruning mechanism of Distribution Tree as specified in [RFC6325] and [RFC7180bis] can also be used for forwarding of multi-destination data frames on the branches that are not pruned.

4.5.2 Unicast Frame processing

Unicast frames must be forwarded in same way they get forwarded in a standard TRILL Campus as specified in [RFC6325]. If multiple equal cost paths are available over pseudowires to reach destination, then VTSD should be capable of doing ECMP for them.

5. Extensions to TRILL Over Pseudowires [RFC7173]

The [RFC7173] mentions how to interconnect a pair of Transparent Interconnection of Lots of Links (TRILL) switch ports using pseudowires. This document explains, how to connect multiple TRILL sites (not limited to only two sites) using the mechanisms and encapsulations defined in [RFC7173].

6. VPTS Model Versus VPLS Model

VPLS Model uses a simpler loop breaking rule: the "split horizon" rule, where a PE must not forward traffic from one PW to another in the same VPLS mesh.

An issue with the above rule is that if a pseudowire between PEs fails, frames will not get forwarded between the PEs where pseudowire went down.

VPTS solves this problem, since the VPTS Model uses distribution Trees for loop free topology, so frames reach all TIRs even when any one of the pseudowires fails in a mesh topology.

If equal cost paths are available to reach a site over pseudowires, VPTS Model can use ECMP for processing of frames over pseudowires.

7. Security Considerations

For general TRILL security considerations, see [RFC6325]

For transport of TRILL by Pseudowires security consideration, see [RFC7173].

Since VPTS Model uses Distribution tree for processing of multi-destination data frames, it is always advisable to run at least one Distribution tree in a TRILL site per tenant, this will avoid data frames getting received on TRILL sites where end-station service is not enabled for that data frame.

8. IANA Considerations

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

9. References

9.1 Normative References

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A.Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RF7180bis] Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., A.Ghanwani, and Gupta, S, "Routing Bridges (RBridges): TRILL: Clarifications, Corrections, and Updates", work in progress. "https://tools.ietf.org/html/draft-ietf-trill-rfc7180bis-05"

[RFC7173] Yong, L., Eastlake 3rd, D., Aldrin, S., and Hudson, J, "Transparent Interconnection of Lots of Links (TRILL) Transport Using Pseudowires", RFC 7173, May 2014.

[RFC4762] Lasserre, M., and Kompella, V., Virtual Private LAN

Service (VPLS) Using Label Distribution Protocol (LDP) Signaling, RFC 4762, January 2007

[RFC4026] Andersson, L., and Madsen, T., Provider Provisioned Virtual Private Network (VPN) Terminology, RFC 4026, March 2005

[RFC4664] Andersson, L., and Rosen, E., Framework for Layer 2 Virtual Private Networks (L2VPNs), RFC 4664, September 2006

9.2 Informative References

[IS-IS] ISO/IEC 10589:2002, Second Edition, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", 2002.

[RFC3985] Bryant, S., Ed., and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC7177] Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014.

[RFC7172] Eastlake 3rd, D., Zhang, R., Agarwal, P., Perlman, R., and Dutt, D, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014.

Authors' Addresses

Mohammed Umair
IP Infusion
RMZ Centennial
Mahadevapura Post
Bangalore - 560048 India

EMail: mohammed.umair2@gmail.com

Kingston Smiler
IP Infusion
RMZ Centennial
Mahadevapura Post
Bangalore - 560048 India

EMail: kingstonsmiler@gmail.com

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

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

Lucy Yong
Huawei Technologies
5340 Legacy Drive
Plano, TX 75024

USA

Phone: +1-469-227-5837
EMail: lucy.yong@huawei.com

TRILL Working Group
INTERNET-DRAFT
Intended status: Informational

Radia Perlman
EMC
Donald Eastlake
Mingui Zhang
Huawei
Anoop Ghanwani
Dell
Hongjun Zhai
JIT

Expires: January 4, 2016

July 5, 2015

Alternatives for Multilevel TRILL (Transparent Interconnection of Lots of Links) <draft-perlman-trill-rbridge-multilevel-10.txt>

Abstract

Extending TRILL to multiple levels has challenges that are not addressed by the already-existing capability of IS-IS to have multiple levels. One issue is with the handling of multi-destination packet distribution trees. Another issue is with TRILL switch nicknames. There have been two proposed approaches. One approach, which we refer to as the "unique nickname" approach, gives unique nicknames to all the TRILL switches in the multilevel campus, either by having the level-1/level-2 border TRILL switches advertise which nicknames are not available for assignment in the area, or by partitioning the 16-bit nickname into an "area" field and a "nickname inside the area" field. The other approach, which we refer to as the "aggregated nickname" approach, involves hiding the nicknames within areas, allowing nicknames to be reused in different areas, by having the border TRILL switches rewrite the nickname fields when entering or leaving an area. Each of those approaches has advantages and disadvantages. This informational document suggests allowing a choice of approach in each area. This allows the simplicity of the unique nickname approach in installations in which there is no danger of running out of nicknames and allows the complexity of hiding the nicknames in an area to be phased into larger installations on a per-area basis.

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79. Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list <trill@ietf.org>.

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

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lot of Links or Tunneled Routing in the Link Layer) protocol [RFC6325] [RFC7177] provides optimal pair-wise data routing without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic in networks with arbitrary topology and link technology, including multi-access links. TRILL accomplishes this by using IS-IS (Intermediate System to Intermediate System [IS-IS] [RFC7176]) link state routing in conjunction with a header that includes a hop count. The design supports data labels (VLANs and Fine Grained Labels [RFC7172]) and optimization of the distribution of multi-destination data based on VLANs and multicast groups. Devices that implement TRILL are called TRILL Switches or RBridges.

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

1.1 TRILL Scalability Issues

There are multiple issues that might limit the scalability of a TRILL-based network:

1. the routing computation load,
2. the volatility of the link state database (LSDB) creating too much control traffic,
3. the volatility of the LSDB causing the TRILL network to be in an unconverged state too much of the time,
4. the size of the LSDB,
5. the limit of the number of TRILL switches, due to the 16-bit nickname space,
6. the traffic due to upper layer protocols use of broadcast and multicast, and
7. the size of the end node learning table (the table that remembers (egress TRILL switch, label/MAC) pairs).

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

IS-IS was designed to be multilevel [IS-IS]. A network can be partitioned into "areas". Routing within an area is known as "Level 1 routing". Routing between areas is known as "Level 2 routing". The Level 2 IS-IS network consists of Level 2 routers and links between the Level 2 routers. Level 2 routers may participate in one or more Level 1 areas, in addition to their role as Level 2 routers.

Each area is connected to Level 2 through one or more "border

routers", which participate both as a router inside the area, and as a router inside the Level 2 "area". Care must be taken that it is clear, when transitioning multi-destination packets between Level 2 and a Level 1 area in either direction, that exactly one border TRILL switch will transition a particular data packet between the levels or else duplication or loss of traffic can occur.

1.2 Improvements Due to Multilevel

Partitioning the network into areas solves the first four scalability issues described above, namely,

1. the routing computation load,

2. the volatility of the LSDB creating too much control traffic,

3. the volatility of the LSDB causing the TRILL network to be in an unconverged state too much of the time,

4. the size of the LSDB.

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

Problem #5 in Section 1.1, namely, the limit of the number of TRILL switches, due to the 16-bit nickname space, will only be addressed with the aggregated nickname approach. Since the aggregated nickname approach requires some complexity in the border TRILL switches (for rewriting the nicknames in the TRILL header), the design in this document allows a campus with a mixture of unique-nickname areas, and aggregated-nickname areas. Nicknames must be unique across all Level 2 and unique-nickname area TRILL switches, whereas nicknames inside an aggregated-nickname area are visible only inside the area. Nicknames inside an aggregated-nickname area must not conflict with nicknames visible in Level 2 (which includes all nicknames inside unique nickname areas), but the nicknames inside an aggregated-nickname area may be the same as nicknames used within other aggregated-nickname areas.

TRILL switches within an area need not be aware of whether they are in an aggregated nickname area or a unique nickname area. The border TRILL switches in area A1 will claim, in their LSP inside area A1, which nicknames (or nickname ranges) are not available for choosing as nicknames by area A1 TRILL switches.

1.3 Unique and Aggregated Nicknames

We describe two alternatives for hierarchical or multilevel TRILL. One we call the "unique nickname" alternative. The other we call the "aggregated nickname" alternative. In the aggregated nickname alternative, border TRILL switches replace either the ingress or egress nickname field in the TRILL header of unicast packets with an aggregated nickname representing an entire area.

The unique nickname alternative has the advantage that border TRILL switches are simpler and do not need to do TRILL Header nickname modification. It also simplifies testing and maintenance operations that originate in one area and terminate in a different area.

The aggregated nickname alternative has the following advantages:

- it solves problem #5 above, the 16-bit nickname limit, in a simple way,
- it lessens the amount of inter-area routing information that must be passed in IS-IS, and
- it logically reduces the RPF (Reverse Path Forwarding) Check information (since only the area nickname needs to appear, rather than all the ingress TRILL switches in that area).

In both cases, it is possible and advantageous to compute multi-destination data packet distribution trees such that the portion computed within a given area is rooted within that area.

1.3 More on Areas

Each area is configured with an "area address", which is advertised in IS-IS messages, so as to avoid accidentally interconnecting areas. Although the area address had other purposes in CLNP (Connectionless Network Layer Protocol, IS-IS was originally designed for CLNP/DECnet), for TRILL the only purpose of the area address would be to avoid accidentally interconnecting areas.

Currently, the TRILL specification says that the area address must be zero. If we change the specification so that the area address value of zero is just a default, then most of IS-IS multilevel machinery works as originally designed. However, there are TRILL-specific issues, which we address below in this document.

1.4 Terminology and Acronyms

This document generally uses the acronyms defined in [RFC6325] plus the additional acronym DBRB. However, for ease of reference, most acronyms used are listed here:

CLNP - ConnectionLess Network Protocol

DECnet - a proprietary routing protocol that was used by Digital Equipment Corporation. "DECnet Phase 5" was the origin of IS-IS.

Data Label - VLAN or Fine Grained Label [RFC7172]

DBRB - Designated Border RBridge

ESADI - End Station Address Distribution Information

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

LSDB - Link State Data Base

LSP - Link State PDU

PDU - Protocol Data Unit

RBridge - Routing Bridge, an alterntive name for a TRILL switch

RPF - Reverse Path Forwarding

TLV - Type Length Value

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

TRILL switch - a device that implements the TRILL protcol [RFC6325], sometimes called an RBridge

VLAN - Virtual Local Area Network

2. Multilevel TRILL Issues

The TRILL-specific issues introduced by multilevel include the following:

a. Configuration of non-zero area addresses, encoding them in IS-IS PDUs, and possibly interworking with old TRILL switches that do not understand nonzero area addresses.

   See Section 2.1.

b. Nickname management.

   See Sections 2.5 and 2.2.

c. Advertisement of pruning information (Data Label reachability, IP multicast addresses) across areas.

   Distribution tree pruning information is only an optimization, as long as multi-destination packets are not prematurely pruned. For instance, border TRILL switches could advertise they can reach all possible Data Labels, and have an IP multicast router attached. This would cause all multi-destination traffic to be transmitted to border TRILL switches, and possibly pruned there, when the traffic could have been pruned earlier based on Data Label or multicast group if border TRILL switches advertised more detailed Data Label and/or multicast listener and multicast router attachment information.

d. Computation of distribution trees across areas for multi-destination data.

   See Section 2.3.

e. Computation of RPF information for those distribution trees.

   See Section 2.4.

f. Computation of pruning information across areas.

   See Sections 2.3 and 2.6.

g. Compatibility, as much as practical, with existing, unmodified TRILL switches.

   The most important form of compatibility is with existing TRILL fast path hardware. Changes that require upgrade to the slow path firmware/software are more tolerable. Compatibility for the relatively small number of border TRILL switches is less important than compatibility for non-border TRILL switches.

See Section 5.

2.1 Non-zero Area Addresses

The current TRILL base protocol specification [RFC6325] [RFC7177] [rfc7180bis] says that the area address in IS-IS must be zero. The purpose of the area address is to ensure that different areas are not accidentally merged. Furthermore, zero is an invalid area address for layer 3 IS-IS, so it was chosen as an additional safety mechanism to ensure that layer 3 IS-IS would not be confused with TRILL IS-IS. However, TRILL uses other techniques to avoid such confusion, such as different multicast addresses and Ethertypes on Ethernet [RFC6325], different PPP (Point-to-Point Protocol) codepoints on PPP [RFC6361], and the like, so use in TRILL of an area address that might be used in layer 3 IS-IS is not a problem.

Since current TRILL switches will reject any IS-IS messages with nonzero area addresses, the choices are as follows:

a.1 upgrade all TRILL switches that are to interoperate in a potentially multilevel environment to understand non-zero area addresses,
a.2 neighbors of old TRILL switches must remove the area address from IS-IS messages when talking to an old TRILL switch (which might break IS-IS security and/or cause inadvertent merging of areas),
a.3 ignore the problem of accidentally merging areas entirely, or
a.4 keep the fixed "area address" field as 0 in TRILL, and add a new, optional TLV for "area name" to Hellos that, if present, could be compared, by new TRILL switches, to prevent accidental area merging.

In principal, different solutions could be used in different areas but it would be much simpler to adopt one of these choices uniformly.

2.2 Aggregated versus Unique Nicknames

In the unique nickname alternative, all nicknames across the campus must be unique. In the aggregated nickname alternative, TRILL switch nicknames within an aggregated area are only of local significance, and the only nickname externally (outside that area) visible is the "area nickname" (or nicknames), which aggregates all the internal nicknames.

The unique nickname approach simplifies border TRILL switches.

The aggregated nickname approach eliminates the potential problem of

nickname exhaustion, minimizes the amount of nickname information that would need to be forwarded between areas, minimizes the size of the forwarding table, and simplifies RPF calculation and RPF information.

2.2.1 More Details on Unique Nicknames

With unique cross-area nicknames, it would be intractable to have a flat nickname space with TRILL switches in different areas contending for the same nicknames. Instead, each area would need to be configured with a block of nicknames. Either some TRILL switches would need to announce that all the nicknames other than that block are taken (to prevent the TRILL switches inside the area from choosing nicknames outside the area's nickname block), or a new TLV would be needed to announce the allowable nicknames, and all TRILL switches in the area would need to understand that new TLV. An example of the second approach is given in [NickFlags].

Currently the encoding of nickname information in TLVs is by listing of individual nicknames; this would make it painful for a border TRILL switch to announce into an area that it is holding all other nicknames to limit the nicknames available within that area. The information could be encoded as ranges of nicknames to make this somewhat manageable [NickFlags]; however, a new TLV for announcing nickname ranges would not be intelligible to old TRILL switches.

There is also an issue with the unique nicknames approach in building distribution trees, as follows:

> With unique nicknames in the TRILL campus and TRILL header nicknames not rewritten by the border TRILL switches, there would have to be globally known nicknames for the trees. Suppose there are k trees. For all of the trees with nicknames located outside an area, the local trees would be rooted at a border TRILL switch or switches. Therefore, there would be either no splitting of multi-destination traffic with the area or restricted splitting of multi-destination traffic between trees rooted at a highly restricted set of TRILL switches.
>
> As an alternative, just the "egress nickname" field of multi-destination TRILL Data packets could be mapped at the border, leaving known unicast packets un-mapped. However, this surrenders much of the unique nickname advantage of simpler border TRILL switches.

Scaling to a very large campus with unique nicknames might exhaust the 16-bit TRILL nicknames space. One method might be to expand nicknames to 24 bits; however, that technique would require TRILL

message format changes and that all TRILL switches in the campus understand larger nicknames.

For an example of a more specific multilevel proposal using unique nicknames, see [DraftUnique].

2.2.2 More Details on Aggregated Nicknames

The aggregated nickname approach enables passing far less nickname information. It works as follows, assuming both the source and destination areas are using aggregated nicknames:

There are two ways areas could be identified.

One method would be to assign each area a 16-bit nickname. This would not be the nickname of any actual TRILL switch. Instead, it would be the nickname of the area itself. Border TRILL switches would know the area nickname for their own area(s).

Alternatively, areas could be identified by the set of nicknames the identify the border routers for that area. (See [SingleName] for a multilevel proposal using such a set of nicknames.)

The TRILL Header nickname fields in TRILL Data packets being transported through a multilevel TRILL campus with aggregated nicknames are as follows:

- When both the ingress and egress TRILL switches are in the same area, there need be no change from the existing base TRILL protocol standard in the TRILL Header nickname fields.

- When being transported in Level 2, the ingress nickname is the nickname of the ingress TRILL switch’s area while the egress nickname is either the nickname of the egress TRILL switch’s area or a tree nickname.

- When being transported from Level 1 to Level 2, the ingress nickname is the nickname of the ingress TRILL switch itself while the egress nickname is either a nickname for the area of the egress TRILL switch or a tree nickname.

- When being transported from Level 2 to Level 1, the ingress nickname is a nickname for the ingress TRILL switch’s area while the egress nickname is either the nickname of the egress TRILL switch itself or a tree nickname.

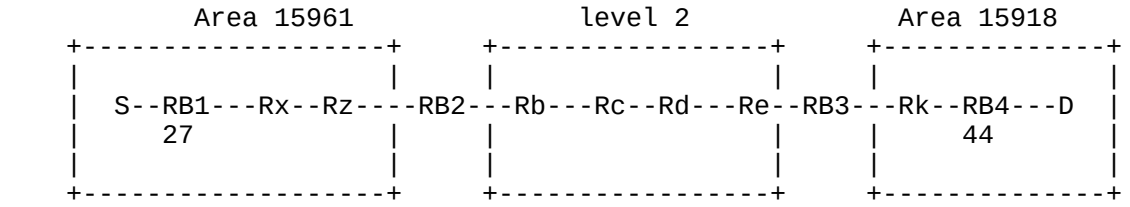
There are two variations of the aggregated nickname approach. The first is the Border Learning approach, which is described in Section

2.2.2.1. The second is the Swap Nickname Field approach, which is described in Section 2.2.2.2. Section 2.2.2.3 compares the advantages and disadvantages of these two variations of the aggregated nickname approach.

2.2.2.1 Border Learning Aggregated Nicknames

This section provides an illustrative example and description of the border learning variation of aggregated nicknames where a single nickname is used to identify an area.

In the following picture, RB2 and RB3 are area border TRILL switches (RBridges). A source S is attached to RB1. The two areas have nicknames 15961 and 15918, respectively. RB1 has a nickname, say 27, and RB4 has a nickname, say 44 (and in fact, they could even have the same nickname, since the TRILL switch nickname will not be visible outside these aggreated areas).



Let’s say that S transmits a frame to destination D, which is connected to RB4, and let’s say that D’s location has already been learned by the relevant TRILL switches. These relevant switches have learned the following:

1) RB1 has learned that D is connected to nickname 15918
2) RB3 has learned that D is attached to nickname 44.

The following sequence of events will occur:

- S transmits an Ethernet frame with source MAC = S and destination MAC = D.

- RB1 encapsulates with a TRILL header with ingress RBridge = 27, and egress = 15918 producing a TRILL Data packet.

- RB2 has announced in the Level 1 IS-IS instance in area 15961, that it is attached to all the area nicknames, including 15918. Therefore, IS-IS routes the packet to RB2. Alternatively, if a distinguished range of nicknames is used for Level 2, Level 1 TRILL switches seeing such an egress nickname will know to route to the nearest border router, which can be indicated by the IS-IS

attached bit.

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with the area nickname, so replaces 27 with 15961. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 15961, and the egress RBridge field will be 15918. Also RB2 learns that S is attached to nickname 27 in area 15961 to accommodate return traffic.

- The packet is forwarded through Level 2, to RB3, which has advertised, in Level 2, reachability to the nickname 15918.

- RB3, when forwarding into area 15918, replaces the egress nickname in the TRILL header with RB4's nickname (44). So, within the destination area, the ingress nickname will be 15961 and the egress nickname will be 44.

- RB4, when decapsulating, learns that S is attached to nickname 15961, which is the area nickname of the ingress.

Now suppose that D's location has not been learned by RB1 and/or RB3. What will happen, as it would in TRILL today, is that RB1 will forward the packet as multi-destination, choosing a tree. As the multi-destination packet transitions into Level 2, RB2 replaces the ingress nickname with the area nickname. If RB1 does not know the location of D, the packet must be flooded, subject to possible pruning, in Level 2 and, subject to possible pruning, from Level 2 into every Level 1 area that it reaches on the Level 2 distribution tree.

Now suppose that RB1 has learned the location of D (attached to nickname 15918), but RB3 does not know where D is. In that case, RB3 must turn the packet into a multi-destination packet within area 15918. In this case, care must be taken so that, in case RB3 is not the Designated transitioner between Level 2 and its area for that multi-destination packet, but was on the unicast path, that another border TRILL switch in that area not forward the now multi-destination packet back into Level 2. Therefore, it would be desirable to have a marking, somehow, that indicates the scope of this packet's distribution to be "only this area" (see also Section 4).

In cases where there are multiple transitioners for unicast packets, the border learning mode of operation requires that the address learning between them be shared by some protocol such as running ESADI [RFC7357] for all Data Labels of interest to avoid excessive unknown unicast flooding.

The potential issue described at the end of Section 2.2.1 with trees in the unique nickname alternative is eliminated with aggregated

nicknames. With aggregated nicknames, each border TRILL switch that will transition multi-destination packets can have a mapping between Level 2 tree nicknames and Level 1 tree nicknames. There need not even be agreement about the total number of trees; just that the border TRILL switch have some mapping, and replace the egress TRILL switch nickname (the tree name) when transitioning levels.

2.2.2.2 Swap Nickname Field Aggregated Nicknames

As a variant, two additional fields could exist in TRILL Data packets we call the "ingress swap nickname field" and the "egress swap nickname field". The changes in the example above would be as follows:

- RB1 will have learned the area nickname of D and the TRILL switch nickname of RB4 to which D is attached. In encapsulating a frame to D, it puts an area nickname of D (15918) in the egress nickname field of the TRILL Header and puts a nickname of RB3 (44) in a egress swap nickname field.

- RB2 moves the ingress nickname to the ingress swap nickname field and inserts 15961, an area nickname for S, into the ingress nickname field.

- RB3 swaps the egress nickname and the egress swap nickname fields, which sets the egress nickname to 44.

- RB4 learns the correspondence between the source MAC/VLAN of S and the { ingress nickname, ingress swap nickname field } pair as it decapsulates and egresses the frame.

See [DraftAggregated] for a multilevel proposal using aggregated swap nicknames with a single nickname representing an area.

2.2.2.3 Comparison

The Border Learning variant described in Section 2.2.2.1 above minimizes the change in non-border TRILL switches but imposes the burden on border TRILL switches of learning and doing lookups in all the end station MAC addresses within their area(s) that are used for communication outside the area. This burden could be reduced by decreasing the area size and increasing the number of areas.

The Swap Nickname Field variant described in Section 2.2.2.2 eliminates the extra address learning burden on border TRILL switches but requires more extensive changes to non-border TRILL switches. In

particular they must learn to associate both a TRILL switch nickname and an area nickname with end station MAC/label pairs (except for addresses that are local to their area).

The Swap Nickname Field alternative is more scalable but less backward compatible for non-border TRILL switches. It would be possible for border and other level 2 TRILL switches to support both Border Learning, for support of legacy Level 1 TRILL switches, and Swap Nickname, to support Level 1 TRILL switches that understood the Swap Nickname method.

2.3 Building Multi-Area Trees

It is easy to build a multi-area tree by building a tree in each area separately, (including the Level 2 "area"), and then having only a single border TRILL switch, say RBx, in each area, attach to the Level 2 area. RBx would forward all multi-destination packets between that area and Level 2.

People might find this unacceptable, however, because of the desire to path split (not always sending all multi-destination traffic through the same border TRILL switch).

This is the same issue as with multiple ingress TRILL switches injecting traffic from a pseudonode, and can be solved with the mechanism that was adopted for that purpose: the affinity TLV [DraftCMT]. For each tree in the area, at most one border RB announces itself in an affinity TLV with that tree name.

2.4 The RPF Check for Trees

For multi-destination data originating locally in RBx’s area, computation of the RPF check is done as today. For multi-destination packets originating outside RBx’s area, computation of the RPF check must be done based on which one of the border TRILL switches (say RB1, RB2, or RB3) injected the packet into the area.

A TRILL switch, say RB4, located inside an area, must be able to know which of RB1, RB2, or RB3 transitioned the packet into the area from Level 2. (or into Level 2 from an area).

This could be done based on having the DBRB announce the transitioner assignments to all the TRILL switches in the area, or the Affinity TLV mechanism given in [DraftCMT], or the New Tree Encoding mechanism discussed in Section 4.1.1.

2.5 Area Nickname Acquisition

In the aggregated nickname alternative, each area must acquire a unique area nickname. It is probably simpler to allocate a block of nicknames (say, the top 4000) to be area addresses, and not used by any TRILL switches.

The nicknames used for area identification need to be advertised and acquired through Level 2.

Within an area, all the border TRILL switches can discover each other through the Level 1 link state database, by using the IS-IS attach bit or by explicitly advertising in their LSP "I am a border RBridge".

Of the border TRILL switches, one will have highest priority (say RB7). RB7 can dynamically participate, in Level 2, to acquire a nickname for identifying the area. Alternatively, RB7 could give the area a pseudonode IS-IS ID, such as RB7.5, within Level 2. So an area would appear, in Level 2, as a pseudonode and the pseudonode could participate, in Level 2, to acquire a nickname for the area.

Within Level 2, all the border TRILL switches for an area can advertise reachability to the area, which would mean connectivity to a nickname identifying the area.

2.6 Link State Representation of Areas

Within an area, say area A1, there is an election for the DBRB, (Designated Border RBridge), say RB1. This can be done through LSPs within area A1. The border TRILL switches announce themselves, together with their DBRB priority. (Note that the election of the DBRB cannot be done based on Hello messages, because the border TRILL switches are not necessarily physical neighbors of each other. They can, however, reach each other through connectivity within the area, which is why it will work to find each other through Level 1 LSPs.)

RB1 acquires an area nickname (in the aggregated nickname approach) and may give the area a pseudonode IS-IS ID (just like the DRB would give a pseudonode IS-IS ID to a link) depending on how the area nickname is handled. RB1 advertises, in area A1, an area nickname that RB1 has acquired (and what the pseudonode IS-IS ID for the area is if needed).

Level 1 LSPs (possibly pseudonode) initiated by RB1 for the area include any information external to area A1 that should be input into area A1 (such as nicknames of external areas, or perhaps (in the unique nickname variant) all the nicknames of external TRILL switches

in the TRILL campus and pruning information such as multicast listeners and labels). All the other border TRILL switches for the area announce (in their LSP) attachment to that area.

Within Level 2, RB1 generates a Level 2 LSP on behalf of the area. The same pseudonode ID could be used within Level 1 and Level 2, for the area. (There does not seem any reason why it would be useful for it to be different, but there's also no reason why it would need to be the same). Likewise, all the area A1 border TRILL switches would announce, in their Level 2 LSPs, connection to the area.

3. Area Partition

It is possible for an area to become partitioned, so that there is still a path from one section of the area to the other, but that path is via the Level 2 area.

With multilevel TRILL, an area will naturally break into two areas in this case.

Area addresses might be configured to ensure two areas are not inadvertently connected. Area addresses appears in Hellos and LSPs within the area. If two chunks, connected only via Level 2, were configured with the same area address, this would not cause any problems. (They would just operate as separate Level 1 areas.)

A more serious problem occurs if the Level 2 area is partitioned in such a way that it could be healed by using a path through a Level 1 area. TRILL will not attempt to solve this problem. Within the Level 1 area, a single border RBridge will be the DBRB, and will be in charge of deciding which (single) RBridge will transition any particular multi-destination packets between that area and Level 2. If the Level 2 area is partitioned, this will result in multi-destination data only reaching the portion of the TRILL campus reachable through the partition attached to the TRILL switch that transitions that packet. It will not cause a loop.

4. Multi-Destination Scope

There are at least two reasons it would be desirable to be able to mark a multi-destination packet with a scope that indicates the packet should not exit the area, as follows:

1. To address an issue in the border learning variant of the aggregated nickname alternative, when a unicast packet turns into a multi-destination packet when transitioning from Level 2 to Level 1, as discussed in Section 4.1.

2. To constrain the broadcast domain for certain discovery, directory, or service protocols as discussed in Section 4.2.

Multi-destination packet distribution scope restriction could be done in a number of ways. For example, there could be a flag in the packet that means "for this area only". However, the technique that might require the least change to TRILL switch fast path logic would be to indicate this in the egress nickname that designates the distribution tree being used. There could be two general tree nicknames for each tree, one being for distribution restricted to the area and the other being for multi-area trees. Or there would be a set of N (perhaps 16) special currently reserved nicknames used to specify the N highest priority trees but with the variation that if the special nickname is used for the tree, the packet is not transitioned between areas. Or one or more special trees could be built that were restricted to the local area.

4.1 Unicast to Multi-destination Conversions

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

In this case, the packet must be converted into a multi-destination packet and flooded in the destination area. However, if the border TRILL switch doing the conversion is not the border TRILL switch designated to transition the resulting multi-destination packet, there is the danger that the designated transitioner may pick up the packet and flood it back into Level 2 from which it may be flooded into multiple areas. This danger can be avoided by restricting any multi-destination packet that results from such a conversion to the destination area through a flag in the packet or though distributing it on a tree that is restricted to the area, or other techniques (see Section 4).

Alternatively, a multi-destination packet intended only for the area could be tunneled (within the area) to the RBridge RBx, that is the appointed transitioner for that form of packet (say, based on VLAN or FGL), with instructions that RBx only transmit the packet within the area, and RBx could initiate the multi-destination packet within the area. Since RBx introduced the packet, and is the only one allowed to transition that packet to Level 2, this would accomplish scoping of the packet to within the area. Since this case only occurs in the unusual case when unicast packets need to be turned into multi-destination as described above, the suboptimality of tunneling between the border TRILL switch that receives the unicast packet and the appointed level transitioner for that packet, would not be an issue.

4.1.1 New Tree Encoding

The current encoding, in a TRILL header, of a tree, is of the nickname of the tree root. This requires all 16 bits of the egress nickname field. TRILL could instead, for example, use the bottom 6 bits to encode the tree number (allowing 64 trees), leaving 10 bits to encode information such as:

- scope: a flag indicating whether it should be single area only, or entire campus
- border injector: an indicator of which of the k border TRILL switches injected this packet

If TRILL were to adopt this new encoding, any of the TRILL switches in an edge group could inject a multi-destination packet. This would require all TRILL switches to be changed to understand the new encoding for a tree, and it would require a TLV in the LSP to indicate which number each of the TRILL switches in an edge group would be.

4.2 Selective Broadcast Domain Reduction

There are a number of service, discovery, and directory protocols that, for convenience, are accessed via multicast or broadcast frames. Examples are DHCP, (Dynamic Host Configuration Protocol) the NetBIOS Service Location Protocol, and multicast DNS (Domain Name Service).

Some such protocols provide means to restrict distribution to an IP subnet or equivalent to reduce size of the broadcast domain they are using and then provide a proxy that can be placed in that subnet to use unicast to access a service elsewhere. In cases where a proxy

mechanism is not currently defined, it may be possible to create one that references a central server or cache. With multilevel TRILL, it is possible to construct very large IP subnets that could become saturated with multi-destination traffic of this type unless packets can be further restricted in their distribution. Such restricted distribution can be accomplished for some protocols, say protocol P, in a variety of ways including the following:

- Either (1) at all ingress TRILL switches in an area place all protocol P multi-destination packets on a distribution tree in such a way that the packets are restricted to the area or (2) at all border TRILL switches between that area and Level 2, detect protocol P multi-destination packets and do not transition them.

- Then place one, or a few for redundancy, protocol P proxies inside each area where protocol P may be in use. These proxies unicast protocol P requests or other messages to the actual campus server(s) for P. They also receive unicast responses or other messages from those servers and deliver them within the area via unicast, multicast, or broadcast as appropriate. (Such proxies would not be needed if it was acceptable for all protocol P traffic to be restricted to an area.)

While it might seem logical to connect the campus servers to TRILL switches in Level 2, they could be placed within one or more areas so that, in some cases, those areas might not require a local proxy server.

5. Co-Existence with Old TRILL switches

TRILL switches that are not multilevel aware may have a problem with calculating RPF Check and filtering information, since they would not be aware of the assignment of border TRILL switch transitioning.

A possible solution, as long as any old TRILL switches exist within an area, is to have the border TRILL switches elect a single DBRB (Designated Border RBridge), and have all inter-area traffic go through the DBRB (unicast as well as multi-destination). If that DBRB goes down, a new one will be elected, but at any one time, all inter-area traffic (unicast as well as multi-destination) would go through that one DRBR. However this eliminates load splitting at level transition.

6. Multi-Access Links with End Stations

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

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

Other methods are possible. For example doing the selection of Appointed Forwarders and of the TRILL switch in charge of that selection across all TRILL switches on the link regardless of area. However, a special case would then have to be made in any case for legacy TRILL switches using area number zero.

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

7. Summary

This draft discusses issues and possible approaches to multilevel TRILL. The alternative using aggregated areas has significant advantages in terms of scalability over using campus wide unique nicknames, not just in avoiding nickname exhaustion, but by allowing RPF Checks to be aggregated based on an entire area. However, the alternative of using unique nicknames is simpler and avoids the changes in border TRILL switches required to support aggregated nicknames. It is possible to support both. For example, a TRILL campus could use simpler unique nicknames until scaling begins to cause problems and then start to introduce areas with aggregated nicknames.

Some issues are not difficult, such as dealing with partitioned areas. Other issues are more difficult, especially dealing with old TRILL switches.

8. Security Considerations

This informational document explores alternatives for the use of multilevel IS-IS in TRILL. It does not consider security issues. For general TRILL Security Considerations, see [RFC6325].

9. IANA Considerations

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

Normative References

[IS-IS] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.

[RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

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

[rfc7180bis] - D. Eastlake, M. Zhang, et al, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, work in progress

Informative References

[RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, August 2011.

[RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014

[RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014.

[RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014, <http://www.rfc-editor.org/info/rfc7177>.

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

[DraftAggregated] - Bhargav Bhikkaji, Balaji Venkat Venkataswami, Narayana Perumal Swamy, "Connecting Disparate Data Center/PBB/Campus TRILL sites using BGP", draft-balaji-trill-

over-ip-multi-level, Work In Progress.

[DraftCMT] - Tissa Senevirathne, Janardhanan Pathang, Jon Hudson, "Coordinated Multicast Trees (CMT) for TRILL", draft-tissa-trill-cmt, Work in Progress.

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

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

[SingleName] - Mingui Zhang, et. al, "Single Area Border RBridge Nickname for TRILL Multilevel", draft-zhang-trill-multilevel-single-nickname-00.txt, Work in Progress.

Acknowledgements

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

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

Authors' Addresses

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

EMail: radia@alum.mit.edu

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

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

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

EMail: zhangmingui@huawei.com

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

EMail: anoop@alumni.duke.edu

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

EMail: honjun.zhai@tom.com

Copyright and IPR Provisions

TRILL Working Group
INTERNET-DRAFT
Intended Status: Standard Track

Y. Li
D. Eastlake
H. Chen
Huawei Technologies
D. Kumar
Cisco
S. Gupta
IP Infusion

Expires: January 6, 2016

July 6, 2015

TRILL: Traceable OAM
draft-yizhou-trill-traceable-oam-00

Abstract

TRILL fault management [RFC7455] specifies the messages and operations for OAM in TRILL network. The sender collects the replies for the OAM-relevant request it sent and uses the replies as the indication of the network faults. In certain circumstances the sender needs to collect multiple replies to isolate the fault, e.g. repetitively sending Path Trace Messages (PTM) with increasing value of hop count and collecting the replies on them to figure out the fault point of certain path.

With the increasing deployment of Software Defined Network (SDN), a centralized management server can be used to help with fault management. The server then is responsible to collect OAM messages and analyze them to either isolate the network fault or compile overall OAM information. It naturally uses SDN structure to alleviate the effort of the requester node and provide a centralized solution to produce the operation and management feedback of the network.

This document specifies the extensions of TRILL OAM message and the operations about the network nodes and the centralized management server to trace and collect OAM relevant messages for further analysis.

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/1id-abstracts.html

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

Table of Contents

1. Introduction

TRILL fault management [RFC7455] specifies OAM messages, TLVs and operations for fault detection and isolation in TRILL network. Requester collects the replies of the OAM requests it sent and tries to analyze the potential faults from them. For path tracing, the sending RBridge utilizes a Hop Count Expiry approach [RFC7455] which is similar as IP trace-route. The sending RBridge sends multiple requests in sequence with incrementing value of Hop Count field and collects the returning responses to construct the path and isolate the fault.

With the deployment of centralized management server in TRILL network, new requirements and approaches for fault isolation and path tracing are emerging. Figure 1 shows a TRILL network with a centralized management server.

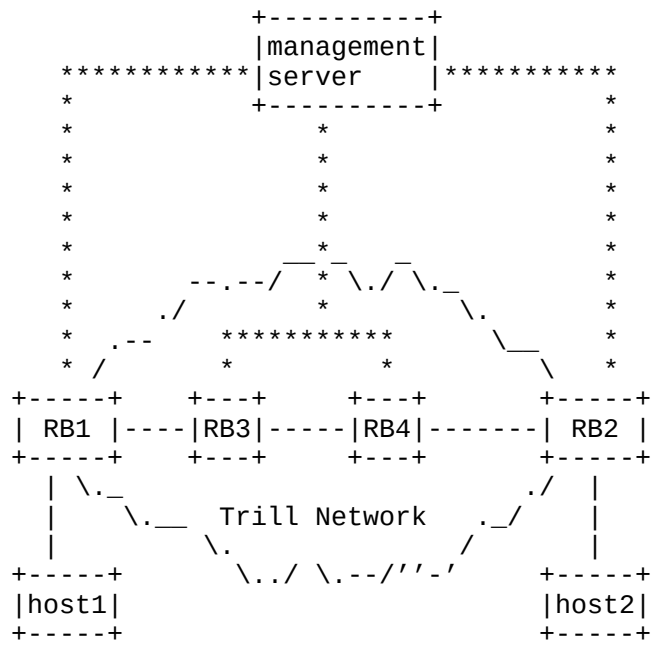


Figure 1. TRILL network with a centralized management server

The centralized management server is capable to construct the OAM messages of a specified flow and feed them into an ingress RBridge, say RB1. When the message is delivered to the egress RBridge RB2, the intermediate RBridges RB3 and RB4 are able to replicate the messages

to the management server. The server is responsible to do all the analysis to trace the path and isolate the fault. Such approach is easily deployable in a network with a controller. For instance, if the management server is an Openflow [Openflow] controller, RBridges may use Packet-in message to send the packets to the Openflow controller and the controller may use Packet-out message to feed the constructed OAM messages into the ingress RB at the beginning.

The document defines the flags and TLVs to help the RBridges to identify the received OAM messages destined for a centralized management server and provides the server with sufficient information for further analysis.

2. Terminology Used in This Document

This document uses the terminology from [RFC6325], [RFC7174] and [RFC7455]. Some additional terms listed below:

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

Data Label: VLAN or FGL.

ECMP: Equal Cost Multi-Path [RFC6325].

FGL: Fine Grained Label [RFC7172].

RBridge: An alternative name for a TRILL switch.

TRILL switch: A device implementing the TRILL protocol. Sometimes called an RBridge.

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

3. Traceable Flag

A new flag ’T’ is defined in TRILL OAM message header [RFC7455] as an indicator for traceable message in figure 2. T flag is applicable to Path Trace Message (PTM) and Multi-Destination Tree Verification Message (MTVM). Loopback message and Continuity Check Message SHOULD not set T flag to 1.

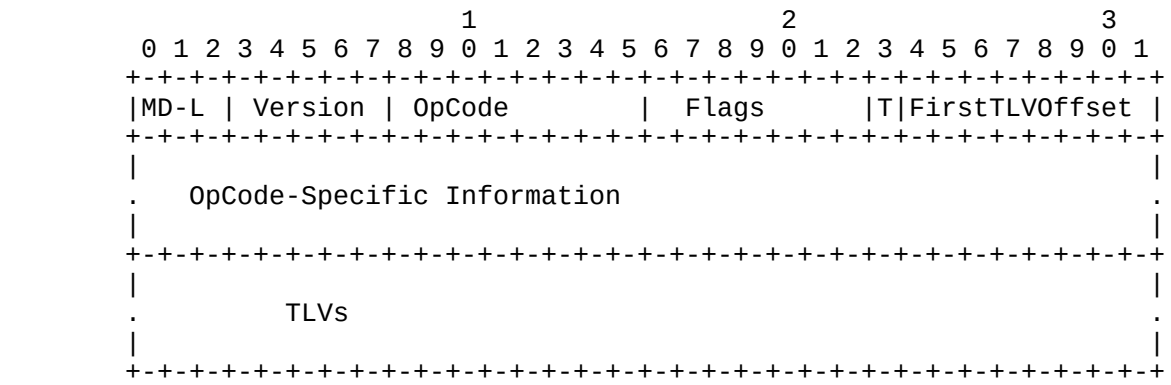


Figure 2. T Flag in TRILL OAM Message Header

o T (1 bit): Traceable flag. When set, indicates no response should be sent back to the requester and the entire TRILL frame should be sent to a centralized management server for tracing.

Basically the traceable flag implies three functions in the trill campus:

1. To indicate the intermediate RBs to capture the frames and replicate it to CPU.

2. To guide the intermediate RB to perform certain operations which may be different from the traditional OAM operations. For example, as we can use packet-in to send the whole packet to openflow controller, it is not necessary to add Original Data Payload TLV to the packet.

3. To make sure the sender will not expect any response and turn off certain mechanisms like time out.

4. Operation Theory

OAM message with Traceable flag is most useful in functionalities requiring tracing, e.g., trace route like behaviors.

4.1 Path Trace Message (PTM) with Traceable Flag

TRILL fault management [RFC7455] adopts an IP trace-route like approach which relies on the hop count expiry to send the PTM message to RBridge for further handling. The sender needs to repetitively send the requests with increasing value of hop count. With traceable flag on, the centralized management server may collect the replicated frame along the path and check the hop count value in TRILL header

directly. By sorting the hop count value decreasingly, it is easy to plot the path taken for a specific flow or figure out the break point for fault isolation.

As a centralized management server normally has more memory space than an RBridge, the server may choose to record the flow entropy to path mapping information. When a fault is suspected between two RBridges, the sever may optimally choose minimum number of flow entropies from the records it saved to feed into the ingress RBridge to spread over the paths.

4.1.1 Actions by Originator RBridge

The originator RBridge takes the following actions:

- Identifies the destination RBridge based on user specification or based on location of the specified destination MAC address.

- Constructs the Flow Entropy based on user-specified parameters or implementation-specific default parameters.

- Specifies the Hop Count of the TRILL Data frame to be larger than the expected number of hops.

- Constructs the TRILL OAM header: set the OpCode to Path Trace Message type (65). Assign an applicable Session Identification number for the request. Return Code and Return Sub-code MUST be set to zero. Set Traceable flags to 1.

- Includes the following OAM TLVs, where applicable:

  - Out-of-Band Reply Address TLV: When Traceable flag is set, Out-of-Band Reply Address TLV needs to be set to the address that the traced message should be sent. It is normally the IP address of the centralized management server. This address may be absent if the default centralized management server address has been configured on every RBridge.

  - Diagnostic Label TLV

  - Sender ID TLV

- Dispatches the OAM frame to the TRILL data plane for transmission.

The originator RBridge SHOULD not expect the replies for the Path Trace Message with Traceable Flag set it sent.

4.1.2 Intermediate RBridge

The intermediate RBridges need to be configured properly as MIP for VLAN/FGL based MA. The TRILL OAM application layer validates the received OAM frame by examining the presence of the TRILL Alert flag, OAM Ethertype at the end of the Flow Entropy, the OpCode being PTM and Traceable Flag set, the intermediate RBridges take the following actions:

- o Optionally include the following TLVs:
  - Previous RBridge Nickname TLV (69)
  - Interface Status TLV (4)
  - Next-Hop RBridge List TLV (70)
  - Sender ID TLV (1)
- o Forward the received message including the TRILL header, the payload and the appended TLVs (if any) to the address specified in Out-of-Band Reply Address TLV. If Out-of-Band Reply Address TLV is not present, either forward it to a system default centralized management server or discard it.

4.1.3 Destination RBridge

Processing is identical to that in Section 4.1.2. The Destination RBridge should not further forward the message in order to prevent leaking of the packet out of the TRILL campus

4.1.4 Centralized Management Server

The centralized management server is normally served as an SDN controller, e.g. an Openflow controller. It is up to the implementation how to deal with the collected packets of PTM with traceable flag from RBridges. The common logic is the centralized management server compares the Session Identification Number and hop count value in TRILL header to trace the path the packet taken.

4.2 Multi-Destination Tree Verification Message (MTVM) with Traceable Flag

MVTM can be used by the OAM tools for plotting the entire or VLAN/FGL pruned distribution tree, reachability verification for set of

RBridges on a given tree or trace along a specified tree to a set of RBridges.

A new TLV is defined as shown in figure 3.

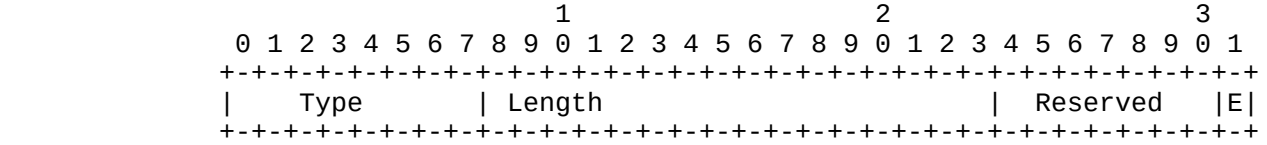


Figure 3. Tree Trace Mode TLV

o Type (1 octet): 75 (TBD), Tree Trace Mode TLV

o Length (2 octets): 1

o E (1 bit): Egress Flag. When RBridge Scope TLV is not present and E flag is 1, trace the receiving RBridge which are egress RBridges on the tree of the specified VLAN/FGL or multicast group; otherwise, ignore this flag.

4.2.1 Actions by Originator RBridge

The originator RBridge takes the following actions:

- Identifies the nickname of distribution tree to be traced.
- Constructs the Flow Entropy based on user-specified parameters or implementation-specific default parameters.
- Specifies the applicable Hop Count value.
- Constructs the TRILL OAM header: set the OpCode to Multicast Tree Verification Message type (67). Assign an applicable Session Identification number for the request. Return Code and Return Sub-code MUST be set to zero. Set Traceable flags to 1.
- Includes the following OAM TLVs, where applicable:
  - Out-of-Band Reply Address TLV: When Traceable flag is set, Out-of-Band Reply Address TLV needs to be set to the address that the traced message should be sent. It is normally the address of the centralized management server
  - RBridge Scope TLV

- Tree Trace Mode TLV: When RBridge Scope TLV is present, E flag of this TLV SHOULD not be set to 1.

- Diagnostic Label TLV

- Sender ID TLV

o Dispatches the OAM frame to the TRILL data plane for transmission.

The originator RBridge SHOULD not expect the replies for the Multicast Tree Verification Message with Traceable Flag it sent.

4.2.2 Receiving RBridge

The TRILL OAM application layer validates the received OAM frame by examining the presence of the TRILL Alert flag and OAM Ethertype at the end of the Flow Entropy. If Traceable Flag is set to 1 in MTVM, the RBridge validates the frame and analyzes if it is an in-scope RBridge.

If the RBridge Scope TLV is present and the local RBridge nickname is specified in the scope list, the receiving RBridge proceeds with further processing as defined in Section 4.1.3.

If the RBridge Scope TLV is absent, the receiving RBridge SHOULD check the Tree Trace Mode TLV. If E flag is 0, the receiving RBridge proceeds with further processing as defined in Section 4.1.3. If E flag is 1 and the receiving RBridge is an egress BRidge for the specified VLAN/FGL or multicast group, the receiving RBridge proceeds with further processing as defined in Section 4.1.3.

For other cases, the receiving RBridge is not considered as in-scope RBridge and should not perform as per section 4.2.3.

4.2.3 In-Scope RBridges

In-Scope RBridges refers to those should tentatively take actions for MTVM request. They are part of receiving RBridges as described in last sub-section.

o Optionally include the following TLVs:

- Previous RBridge Nickname TLV (69)

- Interface Status TLV (4)

- Next-Hop RBridge List TLV (70)

- Sender ID TLV (1)

- Multicast Receiver Port Count TLV (71)

o Forward the received message including the TRILL header, the payload and the appended TLVs (if any) to the address specified in Out-of-Band Reply Address TLV. If Out-of-Band Reply Address TLV is not present, either forward it to a system default centralized management server or discard it.

4.2.4 Centralized Management Server

The centralized management server is normally served as an SDN controller, e.g. an Openflow controller. It is up to the implementation how to deal with the collected packets of MTVM with traceable flag from RBridges. The common logic is the centralized management server compares the Session Identification Number and hop count value in TRILL header to trace the path the packet taken along a distribution tree. It can be used to plot the entire tree or pruned tree or to find out who are the edge RBridges connecting users for a specified VLAN/FGL.

5. Security Considerations

For general TRILL fault management security considerations, please refer to [RFC7455].

6. IANA Considerations

One TLV Type is required to be assigned from the "CFM OAM IETF TLV Types" sub-registry as follows:

Value	Assignment
75	Tree Trace Mode TLV

7. References

7.1 Normative References

[RFC6325] Perlman, R., et.al. "RBridge: Base Protocol Specification", RFC 6325, July 2011.

[RFC6439] Eastlake, D. et.al., "RBridge: Appointed Forwarder", RFC 6439, November 2011.

[RFC6905] Senevirathne, T., Bond, D., Aldrin, S., Li, Y., and R. Watve, "Requirements for Operations, Administration, and Maintenance (OAM) in Transparent Interconnection of Lots of Links (TRILL)", RFC 6905, March 2013.

[RFC7172] Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014.

[RFC7174] Salam, S., Senevirathne, T., Aldrin, S., and D. Eastlake 3rd, "Transparent Interconnection of Lots of Links (TRILL) Operations, Administration, and Maintenance (OAM) Framework", RFC 7174, May 2014,

[RFC7180] Eastlake 3rd, D., Zhang, M., Ghanwani, A., Manral, V., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7180, May 2014.

[RFC7455] Senevirathne, T., Finn, N., Salam, S., Kumar, D., Eastlake 3rd, D., Aldrin, S., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Fault Management", RFC 7455, March 2015.

7.2 Informative References

[OpenFlow] OpenFlow Switch Specification Version, March 26, 2015. (https://www.opennetworking.org/images/stories/downloads/sdn-resources/ onf-specifications/openflow/openflow-switch-v1.5.1.pdf)

8. Acknowledgments

TBD

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56624629
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

Hao Chen
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Email: philips.chenhao@huawei.com

Deepak Kumar
CISCO Systems
510 McCarthy Blvd,
Milpitas, CA 95035, USA

Phone : +1 408-853-9760
Email: dekumar@cisco.com

Sujay Gupta
IP Infusion
RMZ Centennial
Mahadevapura Post
Bangalore - 560048
India

Email: sujay.gupta@ipinfusion.com

INTERNET-DRAFT
Intended Status: Proposed Standard

Mingui Zhang
Donald Eastlake
Huawei
Radia Perlman
EMC
Margaret Wasserman
Painless Security
Hongjun Zhai
JIT

Expires: January 7, 2016
July 6, 2015

Single Area Border RBridge Nickname for TRILL Multilevel
draft-zhang-trill-multilevel-single-nickname-01.txt

Abstract

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

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Copyright and License Notice

Table of Contents

1. Introduction

TRILL multilevel techniques are designed to improve TRILL scalability issues. As described in [MultiL], there have been two proposed approaches. One approach, which is referred as the "unique nickname" approach, gives unique nicknames to all the TRILL switches in the multilevel campus, either by having the Level-1/Level-2 border TRILL switches advertise which nicknames are not available for assignment in the area, or by partitioning the 16-bit nickname into an "area" field and a "nickname inside the area" field. The other approach,

which is referred as the "aggregated nickname" approach, involves assigning nicknames to the areas, and allowing nicknames to be reused in different areas, by having the border TRILL switches rewrite the nickname fields when entering or leaving an area.

The approach specified in this document is different from both "unique nickname" and "aggregated nickname" approach. In this document, the nickname of an area border RBridge is used in both Level 1 (L1) and Level 2 (L2). No additional nicknames are assigned to the L1 areas. Each L1 area is denoted by the group of all nicknames of those border RBridges of the area. For this approach, nicknames in L2 MUST be unique but nicknames inside different L1 areas MAY be reused. The use of the approach specified in this document in one L1 area does not prohibit the use of other approaches in other L1 areas in the same TRILL campus.

2. Acronyms and Terminology

2.1. Acronyms

Data Label: VLAN or FGL

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

2.2. 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].

Familiarity with [RFC6325] is assumed in this document.

3. Nickname Handling on Border RBridges

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

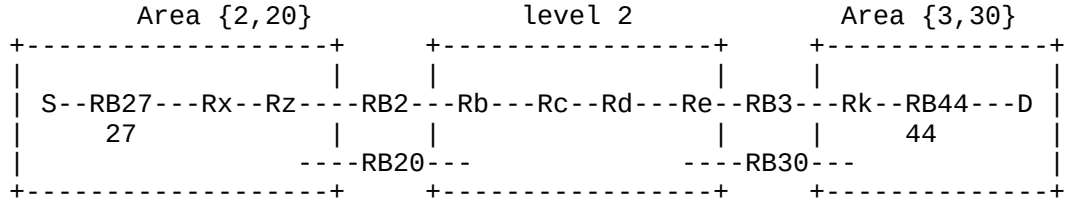


Figure 3.1: An example topology for TRILL multilevel

In Figure 3.1, RB2, RB20, RB3 and RB30 are area border TRILL switches

(RBridges). Their nicknames are 2, 20, 3 and 30 respectively. Area border RBridges use the set of border nicknames to denote the L1 area that they are attached to. For example, RB2 and RB20 use nicknames {2,20} to denote the L1 area on the left.

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

3.1. Actions on Unicast Packets

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

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

The following sequence of events will occur:

- S transmits an Ethernet frame with source MAC = S and destination MAC = D.

- RB27 encapsulates with a TRILL header with ingress RBridge = 27, and egress RBridge = 3 producing a TRILL Data packet.

- RB2 and RB20 have announced in the Level 1 IS-IS instance in area {2,20}, that they are attached to all those area nicknames, including {3,30}. Therefore, IS-IS routes the packet to RB2 (or RB20, if RB20 on the least-cost route from RB27 to RB3).

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with its own nickname, so replaces 27 with 2. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 2, and the egress RBridge field will be 3. (The egress nickname MAY be replaced with an area nickname selected from {3,30}. See Section 4 for the detail of the selection method. Here, suppose nickname 3 is used.) Also RB2 learns that S is attached to nickname 27 in area {2,20} to accommodate return traffic. RB2 SHOULD synchronize with RB20 using ESADI protocol [RFC7357] that MAC = S is attached to nickname 27.

- The packet is forwarded through Level 2, to RB3, which has advertised, in Level 2, its L2 nickname as 3.

- RB3, when forwarding into area {3,30}, replaces the egress nickname in the TRILL header with RB44’s nickname (44). (The

ingress nickname MAY be replaced with an area nickname selected from {2,20}. See Section 4 for the detail of the selection method. Here, suppose nickname 2 is selected.) So, within the destination area, the ingress nickname will be 2 and the egress nickname will be 44.

- RB44, when decapsulating, learns that S is attached to nickname 2, which is one of the area nicknames of the ingress.

3.2. Actions on Multi-Destination Packets

Distribution trees for flooding of multi-destination packets are calculated separately within each L1 area and L2. When a multi-destination packet arrives at the border, it needs to be transitioned either from L1 to L2, or from L2 to L1. All border RBridges are eligible for Level transition. However, for each multi-destination packet, only one of them acts as the Designated Border RBridge (DBRB) to do the transition while other non-DBRBs MUST drop the received copies. All border RBridges of an area SHOULD agree on a pseudorandom algorithm and locally determine the DBRB as they do in the "Per-flow Load Balancing" section. It’s also possible to implement a certain election protocol to elect the DBRB. However, such kind of implementations are out the scope of this document.

As per [RFC6325], multi-destination packets can be classified into three types: unicast packet with unknown destination MAC address (unknown-unicast packet), multicast packet and broadcast packet. Now suppose that D’s location has not been learned by RB27 or the frame received by RB27 is recognized as broadcast or multicast. What will happen, as it would in TRILL today, is that RB27 will forward the packet as multi-destination, setting its M bit to 1 and choosing an L1 tree, flooding the packet on the distribution tree, subject to possible pruning.

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

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with its own nickname, so replaces 27 with 2. RB2 also needs to replace the egress

RBridge nickname with the L2 tree root RBridge nickname, say 2. In order to accommodate return traffic, RB2 records that S is attached to nickname 27 and SHOULD use ESADI protocol to synchronize this attachment information with other border RBridges (say RB20) in the area.

- RB20, will receive the packet flooded on the L2 tree by RB2. It is important that RB20 does not transition this packet back to L1 as it does for a multicast packet normally received from another remote L1 area. RB20 should examine the ingress nickname of this packet. If this nickname is found to be a border RBridge nickname of the area {2,20}, RB2 must not forwarded the packet into this area.

- The packet is flooded on the Level 2 tree to reach both RB3 and RB30. Suppose RB3 is the selected DBRB. The non-DBRB RB30 will drop the packet.

- RB3, when forwarding into area {3,30}, replaces the egress nickname in the TRILL header with the root RBridge nickname, say 3, of the distribution tree of L1 area {3,30}. (Here, the ingress nickname MAY be replaced with an area nickname selected from {2,20} as specified in Section 4.) Now suppose that RB27 has learned the location of D (attached to nickname 3), but RB3 does not know where D is. In that case, RB3 must turn the packet into a multi-destination packet and floods it on the distribution tree of L1 area {3,30}.

- RB30, will receive the packet flooded on the L1 tree by RB3. It is important that RB30 does not transition this packet back to L2. RB30 should also examine the ingress nickname of this packet. If this nickname is found to be an L2 border RBridge nickname, RB30 must not transition the packet back to L2.

- The multicast listener RB44, when decapsulating the received packet, learns that S is attached to nickname 2, which is one of the area nicknames of the ingress.

4. Per-flow Load Balancing

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

4.1. Ingress Nickname Replacement

When a TRILL data packet from other areas arrives at an area border

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

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

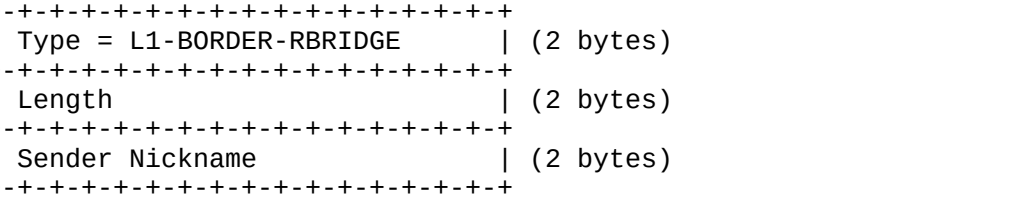
4.2. Egress Nickname Replacement

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

5. Protocol Extensions for Discovery

5.1. Discovery of Border RBridges in L1

The following Level 1 Border RBridge APPsub-TLV will be included in an E-L1FS FS-LSP fragment zero [RFC7180bis] as an APPsub-TLV of the TRILL GENINFO-TLV. Through listening to this Appsub-TLV, an area border RBridge discovers all other area border RBridges in this area.



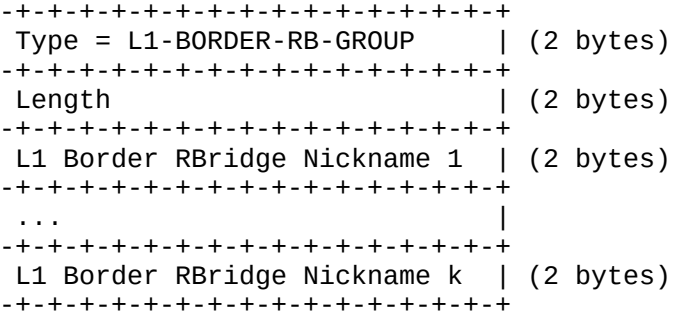
o Type: Level 1 Border RBridge (TRILL APPsub-TLV type tbd1)

o Length: 2

- Sender Nickname: The nickname the originating IS will use as the L1 Border RBridge nickname. This field is useful because the originating IS might own multiple nicknames.

5.2. Discovery of Border RBridge Sets in L2

The following APPsub-TLV will be included in an E-L2FS FS-LSP fragment zero [RFC7180bis] as an APPsub-TLV of the TRILL GENINFO-TLV. Through listening to this APPsub-TLV in L2, an area border RBridge discovers all groups of L1 border RBridges and each such group identifies an area.



- Type: Level 1 Border RBridge Group (TRILL APPsub-TLV type tbd2)

- Length: 2*k. If length is not a multiple of 2, the APPsub-TLV is corrupt and MUST be ignored.

- L1 Border RBridge Nickname: The nickname that an area border RBridge uses as the L1 Border RBridge nickname. The L1-BORDER-RB-GROUP TLV generated by an area border RBridge MUST include all L1 Border RBridge nicknames of the area. It's RECOMMENDED that these k nicknames are ordered in ascending order according to the 2-octet nickname considered as an unsigned integer.

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

6. One Border RBridge Connects Multiple Areas

It's possible that one border RBridge (say RB1) connects multiple L1 areas. RB1 SHOULD use a single area nickname for all these areas.

Nicknames used within one of these areas can be reused within other areas. It's important that packets destined to those duplicated nicknames are sent to the right area. Since these areas are connected to form a layer 2 network, duplicated {MAC, Data Label} across these areas ought not occur. Now suppose a TRILL data packet arrives at the area border nickname of RB1. For a unicast packet, RB1 can lookup the {MAC, Data Label} entry in its MAC table to identify the right destination area (i.e., the outgoing interface) and the egress RBridge's nickname. For a multicast packet: suppose RB1 is not the DBRB, RB1 will not transition the packet; otherwise, RB1 is the DBRB,

- if this packet is originated from an area out of the connected areas, RB1 should replicate this packet and flood it on the proper Level 1 trees of all the areas in which it acts as the DBRB.

- if the packet is originated from one of the connected areas, RB1 should replicate the packet it receives from the Level 1 tree and flood it on other proper Level 1 trees of all the areas in which it acts as the DBRB except the originating area (i.e., the area connected to the incoming interface). RB1 may also receive the replication of the packet from the Level 2 tree. This replication must be dropped by RB1.

7. E-L1FS/E-L2FS Backwards Compatibility

All Level 2 RBridges MUST support E-L2FS [RFC7356] [rfc7180bis]. The Extended TLVs defined in Section 5 are to be used in Extended Level 1/2 Flooding Scope (E-L1FS/E-L2FS) PDUs. Area border RBridges MUST support both E-L1FS and E-L2FS. RBridges that do not support either E-L1FS or E-L2FS cannot serve as area border RBridges but they can well appear in an L1 area acting as non-area-border RBridges.

8. Security Considerations

For general TRILL Security Considerations, see [RFC6325].

The newly defined TRILL APPsub-TLVs in Section 5 are transported in IS-IS PDUs whose authenticity can be enforced using regular IS-IS security mechanism [ISIS][RFC5310]. This document raises no new security issues for IS-IS.

9. IANA Considerations

9.1. TRILL APPsub-TLVs

IANA is requested to allocate two new types under the TRILL GENINFO TLV [RFC7357] for the TRILL APPsub-TLVs defined in Section 5. The following entries are added to the "TRILL APPsub-TLV Types under IS-

IS TLV 251 Application Identifier 1" Registry on the TRILL Parameters IANA web page.

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

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC7356] L. Ginsberg, S. Previdi, et al, "IS-IS Flooding Scope LSPs", RFC 7356, June 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.

10.2. Informative References

[ISIS] ISO, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002.

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.

[RFC7180bis] D. Eastlake, M. Zhang, et al, "TRILL: Clarifications, Corrections, and Updates", draft-eastlake-trill-rfc7180bis, work in progress.

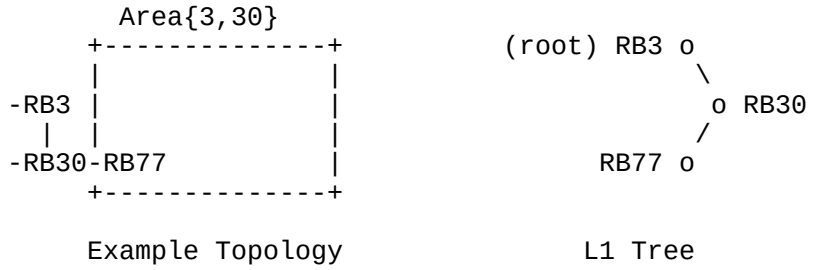
[MultiL] Perlman, R., Eastlake, D., et al, "Flexible Multilevel TRILL", draft-perlman-trill-rbridge-multilevel, work in progress.

Appendix A. Clarifications

A.1. Level Transition

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

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



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

Author's Addresses

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

EMail: zhangmingui@huawei.com

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

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

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

EMail: radia@alum.mit.edu

Margaret Wasserman
Painless Security

EMail: mrw@painless-security.com

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

EMail: honjun.zhai@tom.com