TRILL WG
Internet-Draft
Intended status: Standards Track
Expires: February 18, 2016

Radia. Perlman
EMC Corporation
Fangwei. Hu
ZTE Corporation
Donald. Eastlake 3rd
Huawei technology
Kesava. Krupakaran
Dell
Ting. Liao
ZTE Corporation
August 17, 2015

TRILL Smart Endnodes
draft-ietf-trill-smart-endnodes-02.txt

Abstract

This draft addresses the problem of the size and freshness of the endnode learning table in edge RBridges, by allowing endnodes to volunteer for endnode learning and encapsulation/decapsulation. Such an endnode is known as a "Smart Endnode". Only the attached RBridge can distinguish a "Smart Endnode" from a "normal endnode". The smart endnode uses the nickname of the attached RBridge, so this solution does not consume extra nicknames. The solution also enables Fine Grained Label aware endnodes.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on February 18, 2016.

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] provides optimal pair-wise data frame forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS [IS-IS] [RFC7176] link state routing and encapsulating traffic using a header that includes a hop count. Devices that implement TRILL are called "RBridges" (Routing Bridges) or "TRILL Switches".

An RBridge that attaches to endnodes is called an "edge RBridge" or "edge TRILL Swtich", whereas one that exclusively forwards encapsulated frames is known as a "transit RBridge" or "transit TRILL

Switch". An edge RBridge traditionally is the one that encapsulates a native Ethernet packet with a TRILL header, or that receives a TRILL-encapsulated packet and decapsulates the TRILL header. To encapsulate efficiently, the edge RBridge must keep an "endnode table" consisting of (MAC, Data Label, TRILL egress switch nickname) sets, for those remote MAC addresses in Data Labels currently communicating with endnodes to which the edge RBridge is attached.

These table entries might be configured, received from ESADI [RFC7357], looked up in a directory [RFC7067], or learned from decapsulating received traffic. If the edge RBridge has attached endnodes communicating with many remote endnodes, this table could become large. Also, if one of the MAC addresses and Data Labels in the table has moved to a different remote TRILL switch, it might be difficult for the edge RBridge to notice this quickly, and because the edge RBridge is encapsulting to the incorrect egress RBridge, the traffic will get lost.

2. Solution Overview

The Smart Endnode solution proposed in this document addresses the problem of the size and freshness of the endnode learning table in edge RBridges. An endnode E, attached to an edge RBridge R, tells R that E would like to be a "Smart Endnode", which means that E will encapsulate and decapsulate the TRILL frame, using R’s nickname. Because E uses R’s nickname, this solution does not consume extra nicknames.

Take the below figure as the example Smart Endnode scenario: RB1, RB2 and RB3 are the RBridges in the TRILL domain, and smart SE1 and SE2 are the smart ennodes which can encapsulate and decapsulate the TRILL frames. RB1 is the edge attached RB for SE1 and SE2, and assigns its nickname to SE1 and SE2.

Each Smart Endnode, SE1 and SE2, uses RB1’s nickname when encapsulating, and maintains an endnode table of (MAC, label, TRILL egress switch nickname) for remote endnodes that it (SE1 or SE2) is corresponding with. RB1 does not decapsulate packets destined for SE1 or SE2, and does not learn (MAC, label, TRILL egress switch nickname) for endnodes corresponding with SE1 or SE2, but RB1 does decapsulate, and does learn (MAC, label, TRILL egress switch nickname) for any endnodes attached to RB1 that have not declared themselves to be Smart Endnodes.

Just as an RBridge learns and times out (MAC, label, TRILL egress switch nickname), Smart Endnodes SE1 and SE2 also learn and time out endnode entries. However, SE1 and SE2 might also determine, through ICMP messages or other techniques, that an endnode entry is not

successfully reaching the destination endnode, and can be deleted, even if the entry has not timed out.

If SE1 wishes to correspond with destination MAC D, and no endnode entry exists, SE1 will encapsulate the packet as an unknown destination, or examining updates to the ESADI link state database [RFC7357],or consulting a directory [RFC7067] (just as an RBridge would do if there was no endnode entry).

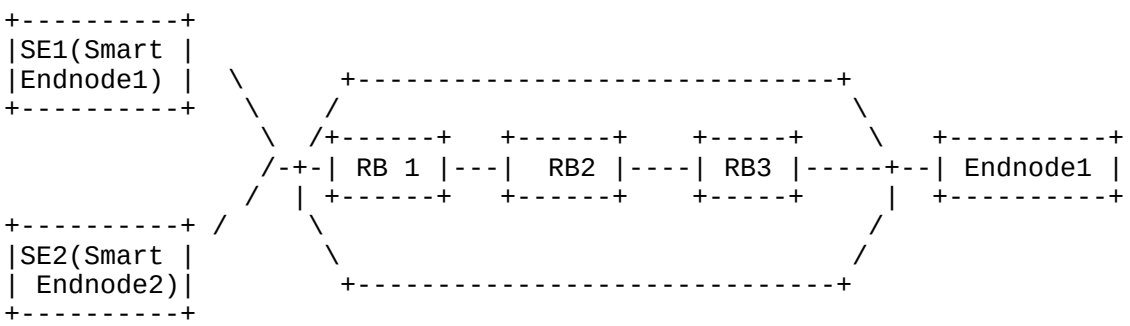

Figure 1 Smart Endnode Scenario

The mechanism in this draft is that the Smart Endnode SE1 issues a Smart-Hello, indicating SE1's desire to act as a Smart Endnode, together with the set of MAC addresses and Data Labels that SE1 owns, and whether SE1 would like to receive ESADI packets. The Smart-Hello is a light type of TRILL-hello, which is used to announce the Smart Endnode capability and parameters (such as MAC address, VLAN ID etc.). The detailed content for a smart endnode's Smart-Hello is defined in section 4.

If RB1 supports having a Smart Endnode neighbor it also sends Smart-Hellos. The smart endnode learns from RB1's Smart-Hellos what RB1's nickname is and which trees RB1 can use when RB1 ingresses multi-destination frames. Although Smart Endnode SE1 transmits Smart-Hellos, it does not transmit or receive LSPs or E-L1FS FS-LSPs[I-D.ietf-trill-rfc7180bis].

Since a Smart Endnode can encapsulate TRILL Data frames, it can cause the Inner.Lable to be a Fine Grained Label [RFC7172], thus this method supports FGL aware endnodes.

3. Terminology

Edge RBridge: An RBridge providing endnode service on at least one of its ports. It is also called an edge TRILL Switch.

Data Label: VLAN or FGL.

ESADI: End Station Address Distribution Information [RFC7357].

FGL: Fine Grained Label [RFC7172].

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

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

Smart Endnode: An endnode that has the capability specified in this document including learning and maintaining(MAC, Data Label, Nickname) entries and encapsulating/decapsulating TRILL frame.

Transit RBridge: An RBridge exclusively forwards encapsulated frames. It is also named as transit RBridge.

TRILL: Transparent Interconnection of Lots of Links [RFC6325].

TRILL switch: a device that implements the TRILL protocol; an alternative term for an RBridge.

4. Smart-Hello Mechanism between Smart Endnode and RBridge

The subsections below describe Smart-Hello messages.

4.1. Smart-Hello Encapsulation

Although a Smart Endnode is not an RBridge, does not send LSPs, and does not perform routing calculations, it is required to have a "Hello" mechanism (1) to announce to edge RBridges that it is a Smart Endnode and (2) to tell them what MAC addresses it is handling in what Data Labels. Similarly, an edge RBridge that supports Smart Endnodes needs a message (1) to announce that support, (2) to inform Smart Endnodes what nickname to use for ingress and what nickname(s) can be used as multi-destination TRILL data packet, and (3) the list of smart end nodes it knows about on that link.

The messages sent by Smart Endnodes and by edge RBridges that support Smart Endnodes are called "Smart-Hellos" and are carried through native RBridge channel messages (see Section 4 of [RFC7178]). They are structured as follows:

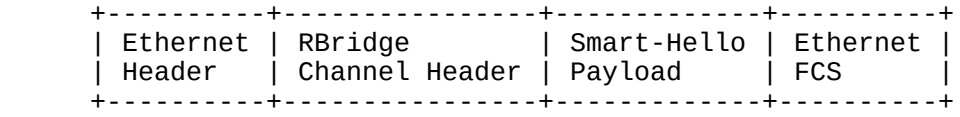

Figure 2 Smart-Hello Structure

In the Ethernet Header, the source MAC address is the address of the Smart Endnode or edge RBridge port on which the message is sent. If the Smart-Hello is sent by a Smart Endnode and multicasted in the link, the destination MAC address is All-Edge-RBridges, and if the Smart-Hello is unicasted to an edge RBridge, the destination MAC address is the MAC address of the RBridge. If the Smart-Hello is sent by an Edge RBridge and multicasted in the link, the destination MAC address is TRILL-End-Stations, and if it is unicasted to a Smart Endnode, the MAC address is the MAC address of the Smart Endnode. The frame is sent in the Designated VLAN of the link so if a VLAN tag is present, it specifies that VLAN.

The RBridge Channel Header begins with the RBridge Channel Ethertype. In the RBridge Channel Header, the Channel Protocol number is as assigned by IANA (see Section 8) and in the flags field, the NA bit is one, the MH bit is zero and the setting of the SL bit is an implementation choice.

The Smart-Hello Payload, both for Smart-Hellos sent by Smart Endnodes and for Smart-Hellos sent by Edge RBridges, consists of TRILL IS-IS TLVs as described in the following two sub-sections. The non-extended format is used so TLVs, sub-TLVs, and APPsub-TLVs have an 8-bit size and type field. Both types of Smart-Hellos MUST include a Smart-Parameters APPsub-TLV as follows inside a TRILL GENINFO TLV:

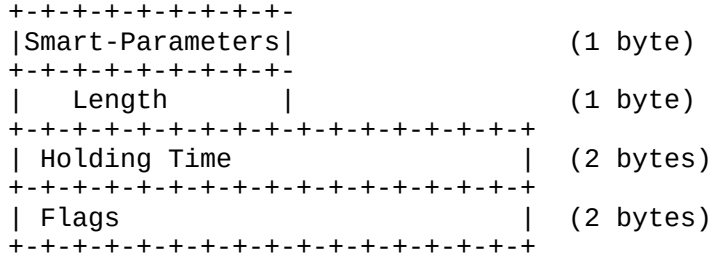

Figure 3 Smart Parameters APPsub-TLV

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

Length: 4.

Holding Time: A time in seconds as an unsigned integer. Has the same meaning as the Holding Time field in IS-IS Hellos [ISIS]. A Smart Endnode and an Edge RBridge supporting Smart Endndoes MUST send a Smart-Hello at least three times during their Holding Time. If no Smart-Hellos is received from a Smart Endnode or Edge RBridge within the most recent Holding Time it sent, it is assumed that it is no longer available.

Flags: At this time all of the Flags are reserved and MUST be send as zero and ignored on receipt.

If more than one Smart Parameters APPsub-TLv appears in a Smart-Hello, the first one is used and any following ones are ignored. If no Smart Parameters APPsub-TLV appears in a Smart-Hello, that Smart-Hello is ignored.

4.2. Edge RBridge's Smart-Hello

The edge RBridge's Smart-Hello contains the following information in addition to the Smart-Parameters APPsub-TLV:

- RBridge's nickname. The nickname sub-TLV (Specified in section 2.3.2 in [RFC7176])is reused here carried inside a TLV 242 (IS-IS router capability) in a Smart-Hello frame. If more than one nickname appears in the Smart-Hello, the first one is used and the following ones are ignored.

- Trees that RB1 can use when ingressing multi-destination frames. The Tree Identifiers Sub-TLV (Specified in section 2.3.4 in [RFC7176]) is reused here.

- Smart Endnode neighbor list. The TRILL Neighbor TLV (Specified in section 2.5 in [RFC7176]) is reused for this purpose.

- An Autentication TLV MAY also be included.

4.3. Smart Endnode's Smart-Hello

A new APPsub-TLV (Smart-MAC TLV) is defined for use by Smart Endnodes as defined below. In addition, there will be a Smart-Parameters APPsub-TLV and there MAY be an Authentication TLV in a Smart Endnode Smart-Hello.

If there are several VLANs/FGL Data Labels for that Smart Endnode, the Smart-MAC APPsub-TLV is included several times in Smart Endnode's Smart-Hello. This APPsub-TLV appears inside a TRILL GENINFO TLV.

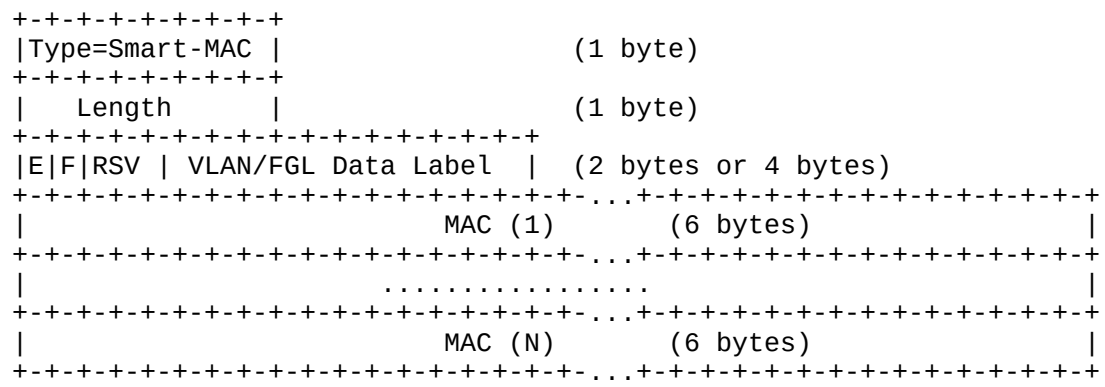

Figure 4 Smart-MAC TLV

- Type: TRILL APPsub-TLV Type Smart-MAC, value is TBD.
- Length: Total number of bytes contained in the value field.
- E: one bit. If it sets to 1, which indicates that the endnode should receive ESADI frames.
- F: one bit. If it sets to 1, which indicates that the endnode supports FGL data label, otherwise, the VLAN/FGL Data Label [RFC7172] field is the VLAN ID.
- RSV: 2 bits or 6 bits, is reserved for the future use. If VLAN/FGL Data Label indicates the VLAN ID(or F flag sets to 0), the RESV field is 2 bits length, otherwise it is 6 bits.
- VLAN/FGL Data Label: This carries a 12-bits VLAN identifier or 24-bits FGL Data Label that is valid for all subsequent MAC addresses in this TLV, or the value zero if no VLAN/FGL data label is specified.
- MAC(i): This is the 48-bit MAC address reachable in the Data Label given from the IS that is announcing this TLV.

5. Data Packet Processing

The subsections below specify Smart Endnode data packet processing. All TRILL data packets sent to or from Smart Endnodes are sent in the Designated VLAN [RFC6325] of the local link but do not necessarily have to be VLAN tagged.

5.1. Data Packet Processing for Smart Endnode

A Smart Endnode does not issue or receive LSPs or E-L1FS FS-LSPs or calculate topology. It does the following:

- o Smart Endnode maintains an endnode table of (the MAC address of remote endnode, Data Label, the nickname of the edge RBridge's attached) entries of end nodes with which the Smart Endnode is communicating. Entries in this table are populated the same way that an edge RBridge populates the entries in its table:
  - * learning from (source MAC address ingress nickname) on packets it decapsulates.
  - * from ESADI[RFC7357].
  - * by querying a directory [RFC7067].
  - * by having some entries configured.
- o When Smart Endnode SE1 wishes to transmit to unicast destination remote node D, if (address of remote endnode D, nickname)entry is in SE1's endnode table, SE1 encapsulates with ingress nickname=the nicknamae of the RBridge(RB1), egress nickname as indicated in D's table entry. If D is unknown, D either queries a directory or encapsulates the packet as a multi-destination frame, using one of the trees that RB1 has specified in RB1's Smart-Hello.
- o When SE1 wishes to transmit to a multicast or broadcast destination, SE1 encapsulates the packet using one of the trees that RB1 has specified.

The Smart Endnode SE1 need not send Smart-Hellos as frequently as normal RBridges. These Smart-Hellos could be periodically unicast to the Appointed Forwarder RB1 through native RBridge channel messages. In case RB1 crashes and restarts, or the DRB changes and SE1 receives the Smart-Hello without mentioning SE1, SE1 SHOULD send a Smart-Hello immediately. If RB1 is AF for any of the VLANs that SE1 claims, RB1 MUST list SE1 in its Smart-Hellos as a Smart Endnode neighbor.

5.2. Data Packet Processing for Edge RBridge

The attached edge RBridge processes and forwards the data frame based on the endnode property rather than for encapsulates and forwards the native frame as the traditional RBridges. There are several situations for the edge RBridges:

- If receiving an encapsulated unicast data frame from a port with a smart endnode, with RB1's nickname as ingress, the edge RBridge RB1 forwards the frame to the specified egress nickname, as with any encapsulated frame. However, RB1 MAY filter the encapsulation frame based on the inner source MAC and Data Label as specified for the Smart Endnode. If the MAC (or Data Label) are not among the expected entries of the Smart Endnode, the frame would be dropped by the edge RBridge.

- If receiving an multi-destination TRILL Data packet from a port with a Smart Endnode, RBridge RB1 forwards the TRILL encapsulation to the TRILL campus based on the distribution tree. If there are some normal endnodes (i.e, non-Smart Endnode) attached to the edge RBridge RB1, RB1 decapsulates the frame and sends the native frame to these ports possibly pruned based on multicast listeners, in addition to forwarding the multi-destination TRILL frame to the rest of the campus.

- When RB1 receives a multicast frame from a remote RBridge, and the exit port includes hybrid endnodes(Smart Endnodes and non-Smart Endnodes), it sends two copies of mulicast frames, one as native and the other as TRILL encapsulated frame. When Smart Endnode receives the encapsulated frame, it learns the remote (MAC address, Data Label, Nickname) entry, A Smart Endnodes ignores native data frames. A normal (non-smart) endnode receives the native frame and learns the remote MAC address and ignores the TRILL data packet. This transit solution may bring some complexity for the edge RBridge and waste network bandwidth resource, so avoiding the hybrid endnodes scenario by attaching the Smart Endnodes and non-Smart Endnodes to different ports is RECOMMENDED. Another solution is that if there are one or more endnodes on a link, the non-Smart Endnodes are ignored on a link; but we can configure a port to support mixed links. If RB1 is configured that the link is "Smart Endnode only", then it will only send and receive TRILL-encapsulated frames on that link. If it is configured to "non-smart-endnodes only" on a port, it will only send and receive native frames from that port.

6. Multi-homing Scenario

Multi-homing is a common scenario for the Smart Endnode. The Smart Endnode is on a link attached to the TRILL domain in two places: to edge RBridge RB1 and RB2. Take the figure below as example. The Smart Endnode SE1 is attached to the TRILL domain by RB1 and RB2 separately. Both RB1 and RB2 could assign their nicknames to SE1.

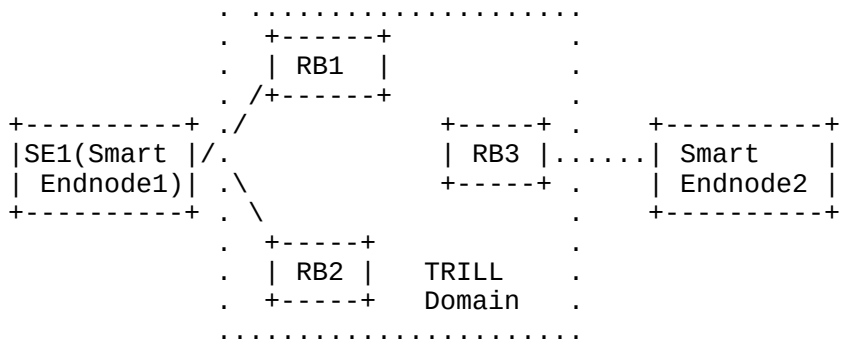

Figure 5 Multi-homing Scenario

There are several solutions for this scenario:

(1) Smart Endnode SE1 can choose either RB1 or RB2's nickname, when encapsulating a frame, whether the encapsulated frame is sent via RB1 or RB2. If SE1 uses RB1's nickname, in this scenario, SE1 will encapsulate with TRILL source nickname RB1 when transmitting on either port. This is simple, but means that all return traffic will be via RB1. If Smart Endnode SE1 wants to do active-active load splitting, and uses RB1's nickname when forwarding through RB1, and RB2's nickname when forwarding through RB2, this will cause MAC flip-flopping of the endnode table entry in the remote RBridges (or Smart Endnodes). One solution is to set a multi- homing bit in the RSV field of the TRILL data packet. When remote RBridge RB3 or Smart Endnodes receives a data packet with the multi-homed bit set, the endnode entries (SE1's MAC addresslabel, RB1's nickname) and (SE1's MAC address, label, RB2's nickname) will coexist as endnode entries in the remote RBridge. Another solution is to extend the ESADI protocol to distribute multiple attachments of a MAC address of a multi-homing group. (Please refer to the option B in section 4 of [I-D.ietf-trill-aa-multi-attach] for details).

(2) RB1 and RB2 might indicate, in their Smart-Hellos, a virtual nickname that attached end nodes may use if they are multihomed to RB1 and RB2, separate from RB1 and RB2's nicknames (which they would also list in their Smart-Hellos). This would be useful if there were many end nodes multihomed to the same set of RBridges. This would be analogous to a pseudonode nickname; return traffic would go via the shortest path from the source to the endnode, whether it is RB1 or RB2. If Smart Endnode SE1 loses connectivity to RB2, then SE1 would revert to using RB1's nickname. In order to avoid RPF check issue for multi-

destination frame, the affinity TLV [I-D.ietf-trill-cmt] is recommended to be used in this solution.

7. Security Considerations

Smart-Hellos can be secured by using Authentication TLVs based on [RFC5310].

For general TRILL Security Considerations, see [RFC6325].

For native RBridge channel Security Considerations, see [RFC7178].

8. IANA Considerations

IANA is requested to allocate an RBridge Channel Protocol number (0x005) to indicate a smart-hello frame.

IANA is requested to allocate APPsub-TLV type numbers for the Smart-MAC and Smart-Parameters APPsub-TLVs.

9. Acknowledgements

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

10. Normative References

[I-D.ietf-trill-aa-multi-attach]
Zhang, M., Perlman, R., Zhai, H., Durrani, M., and S. Gupta, "TRILL Active-Active Edge Using Multiple MAC Attachments", draft-ietf-trill-aa-multi-attach-04 (work in progress), August 2015.

[I-D.ietf-trill-cmt]
Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for TRILL", draft-ietf-trill-cmt-06 (work in progress), March 2015.

[I-D.ietf-trill-rfc7180bis]
Eastlake, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis-05 (work in progress), June 2015.

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

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

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

[RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, DOI 10.17487/RFC7067, November 2013, <http://www.rfc-editor.org/info/rfc7067>.

[RFC7172] Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, DOI 10.17487/RFC7172, May 2014, <http://www.rfc-editor.org/info/rfc7172>.

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

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

[RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357, September 2014, <http://www.rfc-editor.org/info/rfc7357>.

Authors' Addresses

Radia Perlman
EMC Corporation
2010 156th Ave NE, suite #200
Bellevue, WA 98007
USA

Phone: +1-206-291-367
Email: radiaperlman@gmail.com

Fangwei Hu
ZTE Corporation
No.889 Bibo Rd
Shanghai 201203
China

Phone: +86 21 68896273
Email: hu.fangwei@zte.com.cn

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

Phone: +1-508-634-2066
Email: d3e3e3@gmail.com

Kesava Vijaya Krupakaran
Dell
Olympia Technology Park
Guindy Chennai 600 032
India

Phone: +91 44 4220 8496
Email: Kesava_Vijaya_Krupak@Dell.com

Ting Liao
ZTE Corporation
No.50 Ruanjian Ave.
Nanjing, Jiangsu 210012
China

Phone: +86 25 88014227
Email: liao.ting@zte.com.cn