TRILL Working Group
INTERNET-DRAFT
Intended Status: Standard Track

W. Hao
Y. Li
Huawei Technologies
M. Durrani
Cisco
S. Gupta
IP Infusion
A. Qu
MediaTec
T. Han
Huawei Technologies

Expires: September 2016
March 08, 2016

Centralized Replication for Active-Active BUM Traffic
draft-ietf-trill-centralized-replication-05.txt

Abstract

In TRILL active-active access, an RPF check failure issue may occur when using the pseudo-nickname mechanism specified in RFC 7781. This draft describes a solution to resolve this RPF check failure issue through centralized replication. All ingress RBridges send BUM (Broadcast, Unknown unicast and Mutlicast) traffic to a centralized node with unicast TRILL encapsulation. When the centralized node receives the BUM traffic, it decapsulates the packets and forwards them to all destination RBridges using a distribution tree established as per TRILL base protocol RFC 6325. To avoid RPF check failure on a RBridge sitting between the ingress RBridge and the centralized replication node, some change in the RPF calculation algorithm is required. RPF calculation on each RBridge should use the centralized node as the ingress RBridge, instead of the real ingress RBridge, which is denoted as RBv in RFC 7781, to perform the calculation.

Status of this Memo

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

Copyright Notice

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) [RFC6325] protocol provides loop free and per hop based multipath data forwarding with minimum configuration. TRILL uses IS-IS [RFC6165] [RFC7176] as its control plane routing protocol and defines a TRILL specific header for user data.

In active-active, Classic Ethernet (CE) devices typically are multi-homed to edge RBridges which form an edge group. All of the uplinks from CE are handled via a Local Active-Active Link Protocol (LAALP [RFC7379]) such as Multi-Chassis Link Aggregation (MC-LAG) or Distributed Resilient Network Interconnect (DRNI) [802.1AX]. An active-active flow-based load sharing mechanism is normally implemented to achieve better load balancing and high reliability. A CE device can be a layer 3 end system by itself or a bridge switch through which layer 3 end systems access to TRILL campus.

In active-active access, the pseudo-nickname solution in [RFC7781] can be used to avoid MAC flip-flop on remote RBridges. The basic idea is to use a virtual RBridge RBv with a single pseudo-nickname to represent an edge group. Any member RBridge of that edge group MUST use this pseudo-nickname rather than its own nickname as the ingress nickname when it injects TRILL data frames to TRILL campus. The use of the nickname solves the address flip flop issue by binding the MAC address learnt by remote RBridge to the pseudo-nickname. However, it introduces another issue of incorrect packet dropping which will be described as follows: When a pseudo-nickname is used by an edge RBridge as the ingress nickname to forward BUM

traffic, any RBridges (RBn) sitting between the ingress RBridge and the distribution tree root will treat the traffic as if it was ingressed from the virtual RBridge RBv. If the same distribution tree is used by different edge RBridges of the same RBv, the traffic may arrive at RBn from different ports. Then the RPF check fails,

and the BUM traffic received from unexpected ports will be dropped by RBn.

This document proposes a centralized replication solution for broadcast, unknown unicast and multicast (BUM) traffic forwarding to resolve the issue of incorrect packet drop caused by RPF check failure in the virtual RBridge case. The basic idea is that all ingress RBridges send BUM traffic to a centralized node, that SHOULD be a distribution tree root, using unicast TRILL encapsulation. When the centralized node receives the packets, it decapsulates and forwards them to all destination RBridges using a distribution tree established as per the TRILL base protocol.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].The acronyms and terminology in [RFC6325] is used herein with the following additions:

BUM - Broadcast, Unknown unicast and Multicast

CE - As in [RFC7783], Classic Ethernet device (end station or bridge). The device can be either physical or virtual equipment.

FGL - Fine Grained Label [RFC7172].

LAALP - Local Active-Active Link Protocol [RFC7379].

MC-LAG - Multi-Chassis Link Aggregation.

3. Centralized Replication Solution Overview

When an edge RBridge receives BUM traffic from a CE device, it uses unicast TRILL encapsulation instead of multicast encapsulation to send the packets to a centralized node. The centralized node SHOULD be a distribution tree root.

The TRILL header of the unicast TRILL encapsulation contains an "ingress RBridge nickname" field and an "egress RBridge nickname" field. If the ingress RBridge receives the BUM packet from a port which is in an active-active edge group, it should set the ingress RBridge nickname to be the pseudo-nickname rather than its own nickname to avoid MAC flip-flop on remote RBridges as per [RFC7781].

The egress RBridge nickname is set to the special nickname of the centralized node which is used to differentiate the centralized replication purpose unicast TRILL encapsulation from a normal unicast TRILL encapsulation. The special nickname is called an R-nickname.

When the centralized RBridge receives a unicast TRILL encapsulated packet with its R-nickname as egress nickname, it decapsulates the packet. Then the centralized RBridge replicates and forwards the BUM packet to all destination RBridges using one of the distribution trees established as per TRILL base protocol. It SHOULD use a distribution tree whose tree root is the centralized RBridge itself. When the centralized RBridge forwards the BUM traffic, the ingress nickname remains same as that in the packet it received to ensure that the MAC address learning by all egress RBridges is bound to the pseudo-nickname.

When the replicated packet is forwarded by each RBridge along the distribution tree starting from the centralized node, the RPF check will be performed as per [RFC6325]. For any RBridge sitting between the ingress RBridge and the centralized replication node, the incoming port of such BUM packet should be the centralized node facing port as the multicast traffic always comes from the centralized node in this solution. However the RPF port as the result of distribution tree calculation as per [RFC6325] will be the real ingress RBridge facing port as it uses virtual RBridge as the ingress RBridge, so the RPF check will fail. To solve this problem, some change in the RPF calculation algorithm is required. The RPF calculation on each RBridge should use the centralized node as the ingress RBridge instead of the real ingress virtual RBridge to perform the calculation. As a result, RPF check will accept traffic

on the centralized node facing port of the RBridge for multi-destination traffic. This prevents incorrect frame drops by the RPF check.

To differentiate the centralized replication unicast TRILL encapsulation from normal unicast TRILL encapsulation, the R-nickname is introduced for centralized replication. When the centralized node receives unicast TRILL encapsulation traffic with the egress nickname R-nickname, it decapsulates the packet and then forwards the packet to all destination RBridges through a distribution tree by re-encapsulation as aforementioned. The campus through the TRILL LSP extension specified in Section 11.

4. Frame duplication from remote RBridge

Frame duplication may occur when a remote host sends a multi-destination frame to a local CE which has an active-active connection to the TRILL campus. To avoid local CE receiving multiple copies from a remote RBridge, the designated forwarder (DF) mechanism is supported for egress direction multicast traffic.

The DF election mechanism [RFC7781] allows only one port of one RBridge in an active-active group to forward multicast traffic from the TRILL campus to the local access side for each VLAN. The basic idea of DF is to elect one RBridge per VLAN from an edge group to be responsible for egressing the BUM traffic. [RFC7781] describes the detailed DF election mechanism among member RBridges involving in an edge group.

If the DF election mechanism is used for frame duplication prevention, access ports on an RBridge are categorized as three types: non-group, group DF port and group non-DF port. The last two types can be called group ports. Each of the group ports is associated with a pseudo-nickname. If consistent nickname allocation to edge group RBridges is used, it is possible that same pseudo-nickname is associated with more than one port on a single RBridge. A typical scenario is that CE1 is connected to RB1 & RB2 by LAALP1 while CE2 is connected to RB1 & RB2 by LAALP2. In order to conserve the number of pseudo-nicknames used, member ports for both LAALP1

and LAALP2 on RB1 & RB2 are all associated with the same pseudo-nickname.

5. Local forwarding behavior on ingress RBridge

When an ingress RBridge (RB1) receives BUM traffic from a local active-active accessing CE (CE1) device, the traffic will be injected into the TRILL campus with TRILL encapsulation, and it will be replicated and forwarded to all destination RBridges through central replication, including the ingress RBridge itself, along a TRILL distribution tree. To avoid the traffic looping back to the original sender CE, an ingress nickname of the CE group's pseudo-nickname can be used for traffic filtering.

However, if there are two CEs, say CE1 and CE2, connecting to the ingress RB1 and each associated with same pseudo-nickname, RB1 needs to locally replicate and forward to CE2, because another copy of the BUM traffic between CE1 and CE2 through TRILL campus will be blocked by the traffic filtering.

If CE1 and CE2 are not associated with same pseudo-nickname, the copy of the BUM traffic between CE1 and CE2 through TRILL campus won't be blocked by the traffic filtering. To avoid duplicated traffic on receiver CE, there should be no local replicated BUM traffic between these two CEs on ingress RB1.

In summary, to ensure correct BUM traffic forwarding behavior for each CE, the local replication behavior on ingress RBridge should be carefully designed as follows:

1. Replicate to the ports associated with the same pseudo-nickname as that associated to the incoming port.

2. Do not replicate to active-active group ports associated with different pseudo-nicknames.

3. Do not replicate to non-edge-group ports.

The above local forwarding behavior on the ingress RBridge of RB1

can be called centralized replication local forwarding behavior A.

If ingress RBridge RB1 itself is the centralized replication node, BUM traffic injected by RB1 to the TRILL campus won't loop back to RB1. In this case, the local forwarding behavior is called centralized replication local forwarding behavior B. Behavior B on RB1 is as follows:

1. Local replication to the ports associated with the same pseudo-nickname as that associated to the incoming port.

2. Local replication to the group DF port associated with different pseudo-nicknames. Do not replicate to group non-DF port associated with different pseudo-nicknames.

3. Local replication to non-edge-group ports.

6. Loop prevention among RBridges in a edge group

If a CE sends a broadcast, unknown unicast, or multicast (BUM) packet through a DF port to an ingress RBridge, that RBridge will forward that packet to all or a subset of the other RBridges that only have non-DF ports for that active-active group. Because BUM traffic forwarding to non-DF ports isn't allowed, in this case the frame won't loop back to the CE.

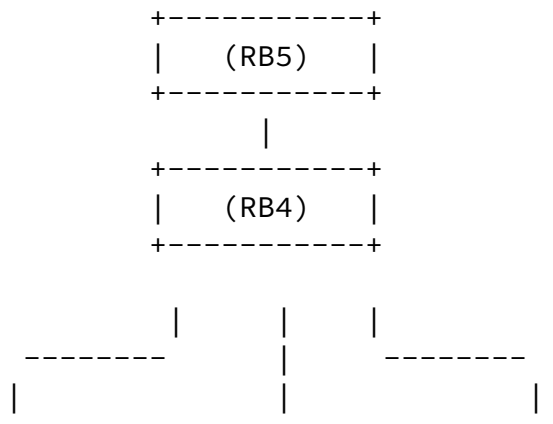
If a CE sends a BUM packet through a non-DF port to a ingress RBridge, say RB1, then RB1 will forward that packet to other RBridges that have a DF port for that active-active group. In this case the frame will loop back to the CE and the traffic split-horizon filtering mechanism is used to avoid looping back among RBridges in the edge group.

This split-horizon mechanism relies on the ingress nickname to check if a packet's egress port belongs to a same active-active group as the packet's incoming port to the TRILL campus.

When the ingress RBridge receives BUM traffic from an active-active accessing CE device, the traffic will be injected into the TRILL campus with TRILL encapsulation, and it will be replicated and

forwarded to all destination RBridges, which include ingress RBridge itself, through a TRILL distribution tree. If the same pseudo-nickname is used for two active-active access CEs as ingress nickname, an egress RBridge can use that nickname to filter traffic forwarding to all local CEs. In this case, the traffic between these two CEs goes through the local RBridge and another copy of the traffic from the TRILL campus is filtered. If different ingress nicknames are used for two connecting CE devices, the access ports connecting to these two CEs should be isolated from each other. The BUM traffic between these two CEs should go through the TRILL campus, otherwise the destination CE connected to same RBridge with the sender CE will receive two copies of the traffic.

7. Centralized replication forwarding process



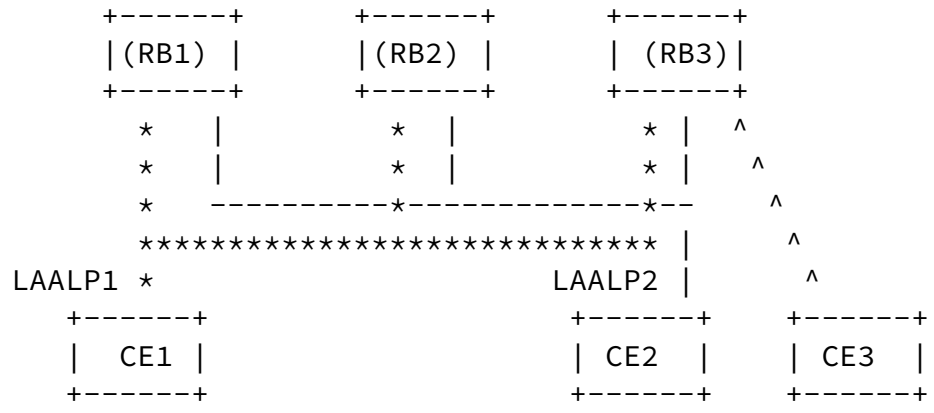


Figure 1 TRILL Active-active access

Assuming the centralized replication solution is used in the example network of above figure 1, RB5 is the distribution tree root and centralized replication node, CE1 and CE2 are active-active accessed to RB1,RB2 and RB3 through LAALP1 and LAALP2 respectively, CE3 is single homed to RB3. The RBridge's own nickname of RB1 to RB5 are nick1 to nick5 respectively. RB1, RB2, and RB3 use the same pseudo-nickname for LAALP1 and LAALP2; that pseudo-nickname is P-nick. The R-nickname on the centralized replication node of RB5 is S-nick.

The BUM traffic forwarding process from CE1 to CE2 and CE3 is as follows:

1. CE1 sends BUM traffic to RB3.

2. RB3 replicates and sends the BUM traffic to CE2 locally. RB2 also sends the traffic to RB5 using unicast TRILL encapsulation. In the TRILL Header, the ingress nickname is set as P-nick and the egress nickname is set as S-nick.

3. RB5 decapsulates the unicast TRILL Data packet. Then it uses the distribution tree whose root is RB5 to forward the packet as a multi-destination TRILL Data packet. The egress nickname in the

multi-destination TRILL Header is the nick5and the ingress nickname is still P-nick.

4. RB4 receives multicast TRILL traffic from RB5. Traffic incoming port is the up port facing the distribution tree root,

RB4's RPF check will be correct based on the changed RPF port calculation algorithm in this document. After the RPF check is performed, it forwards the traffic to all other egress RBridges(RB1, RB2, and RB3).

5. RB3 receives multicast TRILL traffic from RB4. It decapsulates the multi-destination TRILL Data packet. Because the ingress nickname of P-nick is equivalent to the nickname of local LAALPs connecting to CE1 and CE2, RB3 doesn't forward the traffic to CE1 and CE2 to avoid duplicated frame. RB3 only forwards the packet to CE3.

6. RB1 and RB2 receive multicast TRILL traffic from RB4. The forwarding process is similar to the process on RB3, i.e, because the ingress nickname of P-nick is equivalent to the nickname of the local LAALPs connecting CE1 and CE2, they also don't forward the traffic to local CE1 and CE2.

8. BUM traffic loadbalancing among multiple centralized nodes

To support unicast TRILL encapsulation BUM traffic load balancing, multiple centralized replication nodes can be deployed and the traffic can be load balanced between these nodes based on VLAN or FGL.

Assuming there are k centralized nodes in TRILL campus, each centralized node has a different R-nickname, the VLAN-based (or FGL-based [RFC7172]) load balancing algorithm used by ingress active-active access RBridge is as follows:

1. All R-nicknames are ordered and numbered from 0 to k-1 in ascending order treating the nicknames as unsigned 16-bit integers.

2. For VLAN or FGL ID m, choose the R-nickname whose number equals (m mod k) as egress nickname for BUM traffic unicast TRILL encapsulation.

For examples, there are 3 centralized nodes (CN) each having one R-nickname. The CN nodes will be ordered based on the R-nickname from CN0 to CN2. Assuming there are 5 VLANs from VLAN ID 1 to OD 5

spreading among edge RBridges, the traffic in VLAN 1 will go to CN1, VLAN 2 will go to CN2, and so on.

When an ingress RBridge participating in active-active connection receives BUM traffic from local CE, the RBridge decides which centralized node to send the traffic to based on the VLAN-based load balancing algorithm, thus VLAN/FGL-based load balancing for the BUM traffic can be achieved among multiple centralized nodes.

9. Co-existing with the CMT solution

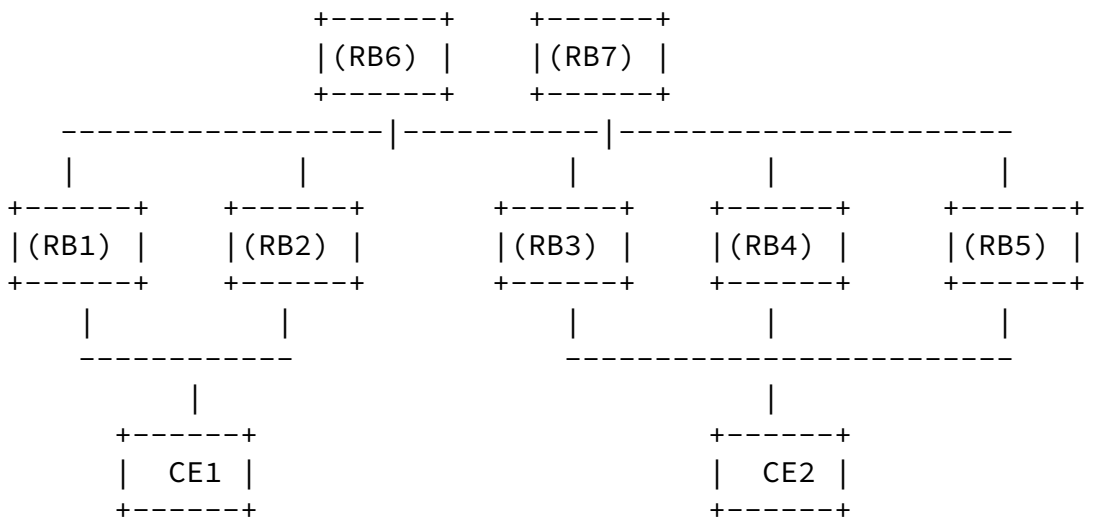


Figure 1 CMT and centralized replication co-existing scenario

Both the centralized replication solution and the CMT [RFC7783] solution rely on using pseudo-nicknames to avoid MAC flip-flop on remote RBridges. These two solutions can co-exist in a single TRILL campus. Each solution can be selected by each active-active edge group of RBridges independently.

As illustrated in figure 2, RB1 and RB2 use CMT for CE1's active-active access, RB3, RB4, and RB5 use the centralized replication for CE2's active-active access.

For the centralized replication solution, edge group RBridges MUST announce the local pseudo-nickname using Nickname Flags APPsub-TLV with C-flag. A nickname with the C-flag set is called a "C-nickname". A transit RBridge will perform the centralized replication specific RPF check algorithm if it receives TRILL Data packets with a C-nickname as ingress nickname.

10. Network Upgrade Analysis

Centralized nodes will typically need software and hardware upgrades to support centralized replication, which stitches TRILL unicast traffic decapsulation process and the process of normal TRILL multicast traffic forwarding along distribution tree.

Active-active connection edge RBridges will typically need software and hardware upgrade to support unicast TRILL encapsulation for BUM traffic; the process is similar to other head-end replication processes.

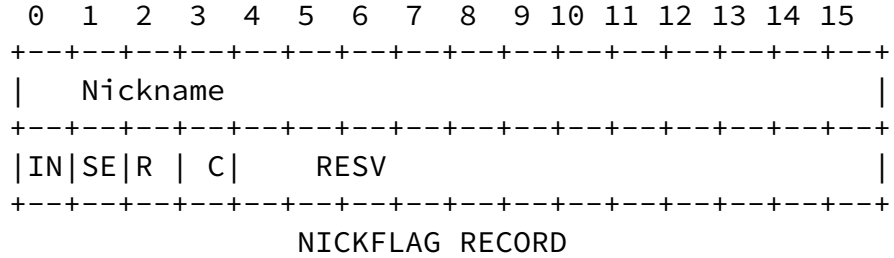
Transit nodes typically need a software upgrade to support the changed RPF port calculation algorithm.

11. TRILL protocol extension

Two Flags of "R" and "C" are specified in the Nickname Flags APPsub-TLV [RFC7780]. The nickname with "R" flag set is called the R-nickname and the nickname with the "C" flag set is called the C-nickname. The R-nickname is a specialized nickname attached on a centralized node to differentiate unicast TRILL encapsulation BUM traffic from normal unicast TRILL traffic. The C-nickname flag is set on each edge group RBridge, C-nickname is a specialized pseudo-nickname for which transit RBridges perform a different RPF check algorithm.

When active-active edge RBridges use centralized replication to forward BUM traffic, the R-nickname is used as the egress nickname and the C-nickname is used as ingress nickname in the TRILL header for the unicast TRILL encapsulation of BUM traffic.

11.1. "R" and "C" Flag in the Nickname Flags APPsub-TLV



NICKFLAG RECORD

o R = If R flag is one, it indicates that the advertising TRILL switch is a centralized replication node, and the nickname is used as egress nickname for edge group RBridges to inject BUM traffic to TRILL campus when the edge group RBridges use centralized

replication solution for active-active access. If flag is zero, that nickname will not be used for that purpose.

o C = If C flag is one, it indicates that the TRILL traffic with this nickname as an ingress nickname requires the special RPF check algorithm. If flag is zero, that nickname will not be used for that purpose.

12. Security Considerations

This draft does not introduce any extra security risks. For general TRILL Security Considerations, see [RFC6325]. For Security Considerations related to pseudo-nickname active-active, see [RFC7781].

13. IANA Considerations

IANA is requested to assign two bits in the Nickname Flags APPsubTLV flags for the R and C bits discussed in Section 11.1 [Bits 3 and 4 suggested] and update the ''NickFlags'' Bits registry on the TRILL Parameters page as follows:

Bit	Mnemonic	Description	Reference
3	R	Replication Nickname	[This document]
4	C	Special RFC Check	[This document]

14. References

14.1. Normative References

[RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, DOI 10.17487/RFC6165, April 2011, <http://www.rfc-editor.org/info/rfc6165>.

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

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

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

[RFC7780] Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <http://www.rfc-editor.org/info/rfc7780>.

14.2. Informative References

[RFC7781] Zhai, H., Senevirathne, T., Perlman, R., Zhang, M., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Pseudo-Nickname for Active-Active Access", RFC 7781, DOI 10.17487/RFC7781, February 2016, <http://www.rfc-editor.org/info/rfc7781>.

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

[RFC7783] Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for Transparent Interconnection of Lots of Links (TRILL)", RFC 7783, DOI 10.17487/RFC7783, February 2016, <http://www.rfc-editor.org/info/rfc7783>.

15. Acknowledgments

The authors wish to acknowledge the important contributions of Donald Eastlake, Hongjun Zhai, Xiaomin Wu, Liang Xia.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: liyizhou@huawei.com

Muhammad Durrani
Cisco
Email: mdurrani@cisco.com

Sujay Gupta
IP Infusion
RMZ Centennial
Mahadevapura Post
Bangalore - 560048
India
Email: sujay.gupta@ipinfusion.com

Andrew Qu
MediaTec
Email: laodulaodu@gmail.com

Tao Han
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: billow.han@huawei.com