TRILL Working Group
Internet Draft
Intended status: Standard Track
Expires: April 2013

L. Dunbar
D. Eastlake
Huawei
Radia Perlman
Intel
Igor Gashinsky
Yahoo
YiZhou Li
Huawei
October 22, 2012

Mechanisms for Directory Assisting TRILL
draft-dunbar-trill-scheme-for-directory-assist-03.txt

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 April 22, 2009.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in

Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

This draft describes the mechanisms of using directory server(s) to assist TRILL (Transparent Interconnection of Lots of Links) edge switches in reducing ARP/ND and unknown unicast flooding across TRILL domain in data center environment.

Conventions used in this document

The term ''Subnet'' and ''VLAN'' are used interchangeably in this document because it is common to map one subnet to one VLAN. The term ''TRILL switch'' and ''RBridge'' are used interchangeably 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].

Table of Contents

1. Introduction 3
2. Terminology 3
3. Push Model of Directory Assisted RBridge Edge in DC Environment 4
   3.1. Minimize the mapping entries maintained by RBridge Edge.. 4
   3.2. Messages to trigger pushing from directory 4
   3.3. Actions by Push Directory Servers 5
   3.4. Applicable Components of ESADI used in Push Scheme..... 5
   3.5. Aggregated entries to push down 6
4. Pull model of Directory Assisted RBridge Edge in DC Environment 7
5. Push-Pull Hybrid Model 9
6. Manageability Considerations 9
7. Security Considerations 9
8. IANA Considerations 9
9. Acknowledgments 10
10. References 10
Authors' Addresses 12
Intellectual Property Statement..... 12
Disclaimer of Validity 13

1. Introduction

[TRILL-Directory-Framework] describes the framework for using directory servers to assist TRILL edge nodes to reduce multi-destination ARP/ND and unknow unicast flooding traffic, thus improving TRILL network scalability in data center environment. This draft describes the detailed mechanisms of using directory servers to assist RBridge edge nodes.

Although this document describes directory servers as being part of RBridges, they may be separate end-stations devices (i.e. standalone directory servers), or co-located with an RBridge.

2. Terminology

AF Appointed Forwarder RBridge port

Bridge: IEEE 802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

DA: Destination Address

DC: Data Center

DDS: Designated Directory Server for a specific VLAN or a group of VLANs

EoR: End of Row switches in data center. Also known as Aggregation switches in some data centers

FDB: Filtering Database for Bridge or Layer 2 switch

Host: Application running on a physical server or a virtual machine. A host usually has at least one IP address and at least one MAC address.

NDSR: Non-Directory Server RBridge. An RBridge which is not directly connected or embedded with a Directory Server

SA: Source Address

STP: Spanning Tree Protocol

RSTP: Rapid Spanning Tree Protocol

ToR: Top of Rack Switch in data center. It is also known as access switches in some data centers.

VM: Virtual Machines

3. Push Model of Directory Assisted RBridge Edge in DC Environment

Under this model, Directory Server(s) push down the IP&MAC&VLAN <-> RBridgeEdge mapping for all the hosts which might communicate with hosts attached to an RBridge edge node. With this environment, it is recommended that RBridge edge simply drop a data packet (instead of flooding to RBridge domain) if the packet's destination address can't be found in the IP&MAC&VLAN<->RBridgeEdge mapping table.

The mapping entry to be pushed down could leverage the gratuitous ARP reply or (Unsolicited) Neighbor Advertisement with extended fields showing the edge RBridge's name, as shown in Table 2.

The push scheme can be accomplished by using the [ESADI] protocol with some simplification or a similar protocol. It is important that it be VLAN scoped.

3.1. Minimize the mapping entries maintained by RBridge Edge

One major drawback of the ''Push Model'' is that RBridge edge's MAC&VLAN<->RBridgeEdge mapping table will have more entries than it really needs.

One simple step for an RBridge to reduce the number of mapping entries pushed down from directory is to prune out entries belonging to VIDs which are not enabled on its bridged LANs ports. For example, if only {vid#1, vid#2, vid#3} are enabled on bridged LANs connected to an RBridge edge ports, only MAC&VLAN<->RBridgeEdge entries for those three VIDs need to be pushed down to the RBridge edge.

To achieve this goal, RBridges need to subscribe directory services for the VLANs which they are interested in. Directory servers only send the directory information to an RBridge for the VLANs subscribed by the RBridge. This process eliminates unnecessary entries to be pushed down to RBridges.

RBridges uses the same mechanism as ESADI protocol to announce all the VLANs which they are interested in since ESADI is already VLAN scoped.

3.2. Messages to trigger pushing from directory

In push down model, it is necessary to have a way for RBridge node to request directory server(s) to start pushing down the mapping entries.

RBridges uses the same mechanism as ESADI protocol to announce, in the IS-IS link state database, all the VLANs which they are interested in. The difference from ESADI is that ''Request for Directory'' message is sent to the Push Directory Servers. All other RBridges who are not attached to any directory servers are not going to process this request.

3.3. Actions by Push Directory Servers

A Push Directory Server could be directly attached to an RBridge or embedded in an RBridge through which VLAN scoped directory contents are pushed to other RBridges.

A Push Directory Server could also be a standalone server which is capable of sending required LSPs to announce its ability and push the content to the subscribed RBridges. A standalone Push Directory is almost like a dummy RBridge node which participates in TRILL link state flooding, but doesn't perform RBridge's forwarding, encapsulating, or decapsulation of native Ethernet data frames.

Push Directory servers advertise their availability by turning on a flag bit in the Interested VLANs sub-TLV [rfc6326bis] in their LSP for the VLAN or VLANs for which they offer Push Directory services. If more than one Directory Server is advertising that it can provide Push Directory Service for a particular VLAN, only the Directory Server associated with the RBridge with the highest System ID on the ESADI pseudo-link [EASDI] should push the information for that VLAN. Other Push Directory servers for that VLAN (presumably present for backup) SHOULD NOT push their directory information to avoid unnecessary duplication.

The Directory Server with the highest System ID is called Designated Directory Server - DDS. Different ESADI pseudo-links [EASDI] for different VLANS could have different DDSs.

There is a reserved Multicast Address for all Push Directory Servers.

3.4. Applicable Components of ESADI used in Push Scheme

RBridges that are not associated with any Push Directory Servers should only participate in ESADI for getting the mapping information for the interested VLAN, but SHOULD NOT advertise any locally learned MAC attachment information into ESADI.

When a non-directory server RBridge detects that the information appears to be missing from the directory information, they can

advertise the information only to the Push Directory Servers by using the well-known multicast address for the Push Directory Servers. This behavior is different from ESADI protocol where the locally learned MAC attachment information is advertised to all RBridges who are interested in the VLANs.

ESADI only advertises the locally learned MAC address. But the directory needs to push down IP/MAC/VLAN and their directly attached RBridges.

[draft-eastlake-isis-ia-tlv] describes a proposed TLV to carry the VLAN scoped IP/MAC/VLAN information for all the hosts.

3.5. Aggregated entries to push down

Using Table 2 requires one entry per host/VM. When directory pushes down the entire mapping to an edge RBridge for the very first time, there usually are many entries. To minimize the amount of data pushed down, summarization should be considered, e.g. with one edge RBridge Nickname being associated with all attached hosts' MAC addresses and VLANs as shown below:

Nickname1	VID-1	MAC1/IP, MAC2/IP, ,, MACn/IP
	VID-2	MAC1/IP, MAC2/IP, ,, MACn/IP
	,,,,,	MAC1/IP, MAC2/IP, ,, MACn/IP
Nickname2	VID-1	MAC1/IP, MAC2/IP, ,, MACn/IP
	VID-2	MAC1/IP, MAC2/IP, ,, MACn/IP
	,,,,,	MAC1/IP, MAC2/IP, ,, MACn/IP
-------	------	MAC1/IP, MAC2/IP, ,, MACn/IP
	,,,,,	MAC1/IP, MAC2/IP, ,, MACn/IP

Table 1: Summarized table pushed down from directory

Whenever there is any change in MAC&VLAN <-> RBridgeEdge mapping, which can be triggered by hosts being added, moved, or de-commissioned, an incremental update can be sent to the RBridge edges which are impacted by the change.

4. Pull model of Directory Assisted RBridge Edge in DC Environment

Under this model, ''RBridge'' pulls the VLAN scoped MAC<->IP<->RBridgeEdge mapping entry from the directory server when needed.

Pull Directory Servers for a particular VLAN are located by looking in the link state database for RBridges that advertise themselves by having the Pull Directory Server flag on in their Interested VLANs sub-TLV [rfc6326bis] for that VLAN. If multiple RBridges indicate that they are Pull Directory Servers for a particular VLAN, then pull requests can be sent to any of them.

Pull Directory requests are sent to the RBridge, or a dummy RBridge, whose LSP contains the Interested VLANs sub-TLV advertising that it is a Pull Directory server for the relevant VLAN. These requests are sent by enclosing them in an RBridge Channel [Channel] message using the Pull Directory channel protocol number (see Section 8). Responses are returned in an RBridge Channel message using the same protocol number.

The requests to Pull Directory Servers are derived from normal ARP [RFC826], ND [RFC4861], RARP [RFC903] messages intercepted by the RBridge, or data frame with unknown DA.

However, additional information is desired from the directory server response in this case, such as the nickname to which an end station (probably identified by IP address) is attached. For this purpose, extended ARP op codes are specified in Table 2.

For a Pull Request derived from an unknown data frame, the RBridge edge node can drop the data frame if there is no response from the directory server after X number of tries.

The requesting RBridge node can cache the mapping and age out MAC&IP&VLAN entries if the entries haven't been used for a certain period of time. Therefore, each RBridge edge will only keep the entries which are frequently used, i.e. mapping table size can be smaller.

The following table shows how target RBridge nickname can be attached to a standard ARP Reply when replying to an ARP request forwarded by ingress RBridge edge.

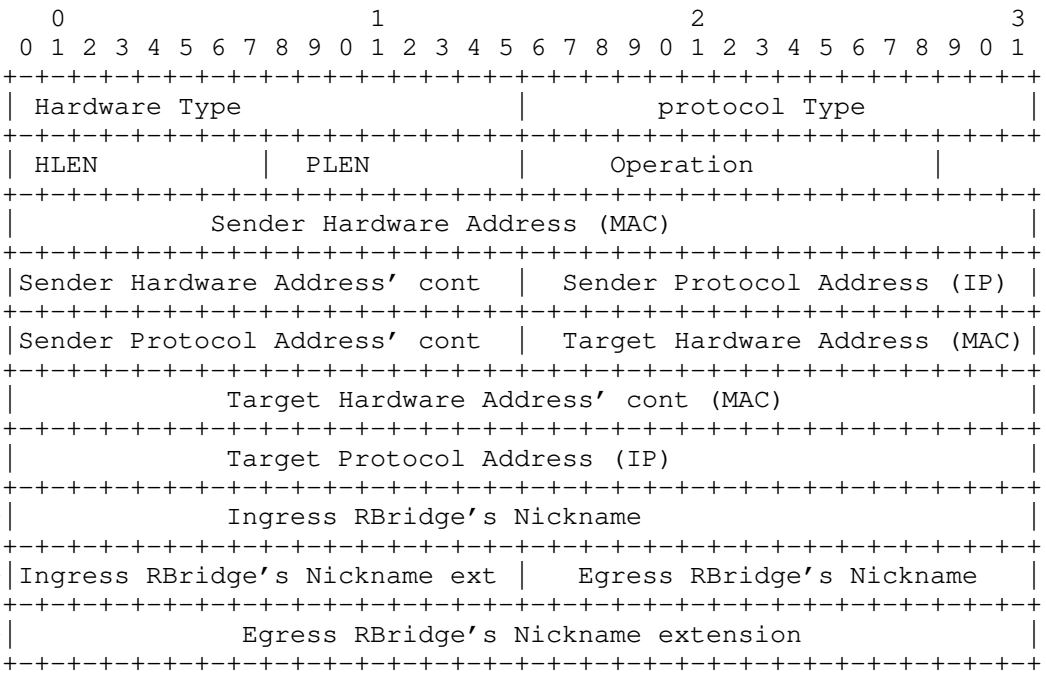


Table 2: Extended fields added to standard ARP reply

The original ARP reply format consists of the first 28 octets shown in this table. The last 12 octets in this table marked by ''->'' are extended fields to indicate the Ingress RBridge to which originating host is attached and the Egress RBridge to which the target host is attached. More bits are reserved for RBridge nicknames in case multiple levels of nicknames are needed in the future for large data centers.

There are 16 bits for Operation type field in ARP message. IANA has assigned 0~25 for various purposes and leave 26~65534 unassigned [http://www.iana.org/assignments/arp-parameters/arp-parameters.xml]. If this approach is taken, a new ARP Operation code has to be assigned by IANA.

It worth noting that the ''Egress RBridge's Nickname in Table 2 is the nickname of the ''Target RBridge'' to which the Target host is attached.

When Pull Directory Server is embedded in an RBridge, the Pull Request would have the ''Ingress RBridge Nickname'' in the TRILL Header. The ''Ingress RBridges Nickname'' field in the Pull Request is for future extension when directory server is not an RBridge.

5. Push-Pull Hybrid Model

For some edge nodes which have great number of VIDs enabled, managing the MAC&VLAN <-> RBridgeEdge mapping for hosts under all those VIDs can be challenge. This is especially true for Data Center gateway nodes, which need to maintain majority of VIDs if not all.

For those RBridge Edge nodes, hybrid model should be considered. I.e. Push model are used for some VIDs, and pull model are used for other VIDs. It can be operator's decision (i.e. by configuration) on which VIDs' mapping entries are pushed down from directory and which VIDs' mapping entries are pulled.

For example, in a data center when hosts in specific VIDs (vid#1, vid#2, ? vid#100)communicate regularly with external peers, the mapping entries for those 100 VIDs should be pushed down to the data center gateway routers. For hosts in other VIDs which only communicate with external peers once a day (or once a few days) for management interface, the mapping entries for those VIDs should be pulled down from directory whenever the needs come up.

The mechanisms described above for Push and Pull Directory services make it easy to use Push for some VIDs and Pull for others. In fact, different RBridges can even be configured so that some use Push Directory services and some use Pull Directory services for the same VID if both Push and Pull Directory services are available for that VID. And there can be VIDs for which directory services are not used.

6. Manageability Considerations

TBD.

7. Security Considerations

For general TRILL security considerations, see [RFC6325].

8. IANA Considerations

There are 16 bits for ARP Operation type field [RFC826]. IANA has assigned 0~25 for various purposes and leave 26~65534 unassigned [http://www.iana.org/assignments/arp-parameters/arp-parameters.xml]. If this approach is taken, IANA is requested to assign a new ARP Operation code for TRILL Directory Pull services.

IANA is request to allocate a new RBridge Channel protocol number for Pull Directory Services.

IANA is requested to allocate two currently reserved bits in the Interested VLANs field of the Interested VLANs and Spanning Tree Roots sub-TLV [rfc6326bis] to indicate directory servers and to create a sub-registry in the TRILL Parameters Registry, as follows:

Interested VLANs Flag Bits

Registration Procedures: IETF Review

Reference:

Bit	Mnemonic	Description	Reference
0	M4	IPv4 Multicast Router Attached	[rfc6326bis]
1	M6	IPv6 Multicast Router Attached	[rfc6326bis]
2	PS	Push Directory Server	This document
3	PL	Pull Directory Server	This document
16-19	-	available for allocation	[rfc6326bis]

9. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

10. References

[TRILL-Directory-Framework] Dunbar, et, al ''TRILL Edge Directory Assistance Framework'', <draft-ietf-trill-directory-framework-02>, work in progress,Oct 2012.

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

[RFC826] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.

[RFC903] Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984

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

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September

[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.

[rfc6326bis] Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, ''TRILL Use of IS-IS'', draft-ietf-isis-rfc6326bis, work in progress

[ESADI] draft-ietf-trill-esadi, work in progress.

[InterfaceAddresses] ''Interface Addresses TLV'', draft-eastlake-isis-ia-tlv, work in progress.

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

[ARMD-Problem] Narten, et,al, ''Problem Statement for ARMD'', June 2012.

[ARP reduction] Shah, et. al., "ARP Broadcast Reduction for Large Data Centers", Oct 2010

Authors' Addresses

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA
Phone: 1-508-333-2270
Email: d3e3e3@gmail.com

Radia Perlman
Intel Labs
2200 Mission College Blvd.
Santa Clara, CA 95054-1549 USA
Phone: +1-408-765-8080
Email: Radia@alum.mit.edu

Igor Gashinsky
Yahoo
45 West 18th Street 6th floor
New York, NY 10011
Email: igor@yahoo-inc.com

YiZhou Li
Huawei
Email: liyizhou@huawei.com

Intellectual Property Statement

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

TRILL Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 27, 2013

H. Zhai
ZTE
T. Senevirathne
Cisco Systems
R. Perlman
Intel Labs
D. Eastlake 3rd
M. Zhang
Huawei
F. Hu
ZTE

August 26, 2012

RBridge: Pseudo-Nickname
draft-hu-trill-pseudonode-nickname-03

Abstract

At the edg of TRILL campus, some RBridges provide end-station services to their attached end stations; these RBridges are called edge RBridges. To avoid potential frame duplication or loops in TRILL campus, only one edge RBridge is permitted to provide such services in a VLAN at all times to its attached end stations even they are also attached to other edge RBridges. However, in some application scenarios, for example in Link Aggregation Group (LAG), more than one edge RBridge is required to provide such services to an end station even in a VLAN to improve resiliency and maximize the available network bandwidth, which causes the flip-flopping of the egress RBridge nickname for such an end station in remote RBridges' forwarding tables. In this document, the concept of Virtual RBridge, along with its pseudo-nickname, is introduced to address the above problem.

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 27, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction 4
1.1. Terminology and Acronyms 5
2. Problem Statement 5
2.1. Appointed Forwarders on Shared Links 5
2.2. Multi-homing and Link Aggregation to TRILL Network 6
3. Concept of Virtual RBridge and Pseudo-nickname 7
3.1. VLAN-x Appointed Forwarder for member interfaces in RBv . 8
3.2. Announcing Pseudo-Nickname of RBv 8
4. Distribution Trees for Member RBridges in RBv 8
5. Frame Processing 10
5.1. Data Frames 10
5.1.1. Native Frames Ingressing 10
5.1.2. TRILL Data Frames Egressing 10
5.1.2.1. Unicast TRILL Data Frames 10
5.1.2.2. Multi-Destination TRILL Data Frames 11
6. Member Link Failure in RBv 12
7. OAM Frames 12
8. Configuration Consistency 13
9. IANA Considerations 13
10. Security Considerations 13
11. Acknowledgements 13
12. Normative References 13
Appendix A. Reasons for MAC Sharing among Member RBriges 14
Authors' Addresses 15

1. Introduction

The IETF TRILL protocol [RFC6325] provides optimal pair-wise data frame forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multi-pathing of both unicast and multicast traffic. TRILL accomplishes this by using [IS-IS] [RFC1195] link state routing and encapsulating traffic using a header that includes a hop count. The design supports VLANs and optimization of the distribution of multi-destination frames based on VLANs and IP derived multicast groups. Devices that implement TRILL are called RBridges.

In TRILL protocol, RBridges are identified by nicknames (16-bits). Different RBridge has different nickname(s). At the edge of TRILL network, some RBridges connect to legacy networks on one side and to TRILL network on the other side. These RBridges are called edge RBridges. For the connectivity between the two types of network, edge RBridges provide end-station services to end stations located in legacy networks. When receiving a native frame from such a local end station S, the service edge RBridge RB1 encapsulates the frame in a TRILL header, addressing the packet to RBridge RBx to which the destination end station D is attached. The TRILL header contains an "ingress RBridge nickname" field (RB1's nickname), an "egress RBridge nickname" field (RBx's nickname), and a hop count. On receiving such a frame, RBx removes the TRILL header and forwards it onto D in its native form. Meanwhile, based on the de-capsulation of such a frame, RBx learns the (ingress RBridge nickname, source MAC address, VLAN ID) triplet. Edge RBridges maintain such triplets in their forwarding table for the potential following transmission of native frames.

Due to failures, reconfiguration and other network dynamics, service edge RBridge for S may change over from RB1 to another edge RBridge. In this event, remote traffic addressed to S will be still forwarded to RB1 by remote RBridge RBx before perceiving this change, and then the traffic gets dropped at RB1, causing traffic disruption. Furthermore, to improve resiliency and maximize the available network bandwidth, an end station typically is multi-homed to several edge RBridges and treats all the uplink links as a Link Aggregation Group (LAG) bundle. In this scenario, all those edge RBridges work in an active-active load sharing model to provide end-station services for an end station even in same VLAN. When remote RBridge RB2 receives different frames, which are originated by such an end station S and ingressed into TRILL campus by different such edge RBridge, flip-flopping of ingress RBridge nickname for MAC of S will be observed by RBx during de-capsulating such frames. This flip-flopping will cause disorder of different frames in traffic, worsening the traffic disruption.

In this document, concept of Virtual RBridge group, together with its Pseudo-nickname, is introduced to address the above issues. For a member RBridge in such a group, it uses the pseudo-nickname of this group, instead of its own device nickname, as ingress RBridge nickname when encapsulating a frame to its TRILL form with a TRILL header. So, in a RBridge Group, even if there are more than one RBridge providing end-station services for a end station or the service RBridge changes over from one member RBridge to another in same set of VLANs, the ingress RBridge nickname for the MAC of this end station will still remain unchanged in remote RBridges' forwarding tables.

This document is organized as following: Section 2 is problem statement, which describes why virtual RBridge and its pseudo-nickname are required. Section 3 gives the concept of virtual RBridge. Section 4 describes the consideration for pseudo-nickname used in ingressing multi-destination frames. Section 5 covers processing of transit frame traffic when considering pseudo-nickname.

Familiarity with [RFC6325] is assumed in this document.

1.1. Terminology and Acronyms

This document uses the acronyms defined in [RFC6325] and the following additional acronym:

AF - Appointed Forwarder

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].

When used in lower case, these words convey their typical use in common language, and are not to be interpreted as described in [RFC2119].

2. Problem Statement

2.1. Appointed Forwarders on Shared Links

If there are multiple RBridges on a shared link, together with end stations, only one RBridge is permitted to provide end-station services in a VLAN at all times for the end stations to avoid possible frame duplication or loops in TRILL campus. The service RBridge is called VLAN-x Appointed Forwarder (AF) on a shared link.

However, AF for any set of VLANs on a shared link may change over

from one RBridge to another, due to failure, configurations and other network dynamics, etc. If such change occurs, local end stations may not perceive it, so the end station cannot timely notify remote RBridges to update the correspondence between ingress RBridge nickname and the MAC of this end station in their forwarding tables. As a result, remote RBridges may continue to forward traffic to the previous AF and the traffic may be dropped at the previous egress RBridge, causing traffic disruption.

2.2. Multi-homing and Link Aggregation to TRILL Network

In order to improve the reliability of connection to TRILL network, multi-homing technique may be employed by a legacy device, such as a switch or end host. For example, in Figure 1, switch SW1 multi-homes to TRILL network by connecting to both RBridge RB1 and RB2 with respective links. Then the end stations (e.g., S1), attached to SW1, still get end-station services from TRILL network even if one connection of SW1 to TRILL network, e.g., SW1-RB1, fails. That is to say, the service RBridge for S1 changes over from RB1 to RB2 after the connection of SW1-RB1 fails, which causes traffic disruption temporarily (e.g., traffic from Sx to S1), similar to AF changing in Section 2.1.

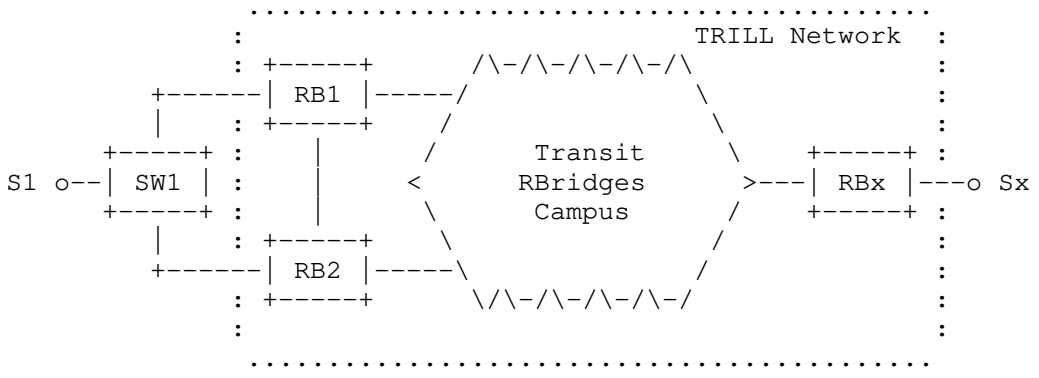


Figure 1
Multi-homing to TRILL Network

Furthermore, SW1 may treat the two links as a LAG (Link Aggregation Group) bundle, so that the two links form active-active load sharing model instead of previous active-stanby model. That is to say, in Figure 1, two service RBridges (e.g., RB1 and RB2) must provide end-station services simultaneously to S1 in that VLAN. And this will result in flip-flopping of the ingress RBridge for the MAC of S1 in remote RBridges' (e.g., RBx) forwarding tables. AS a result, this flip-flopping will cause much more disorder packets and worsen the

traffic disruption.

Besides switches, end stations can also directly multi-home to TRILL network and treat the multi-homing links as a LAG bundle. The issue of traffic disruption also occurs in this scenario if such an end station balances different traffic load in a same VLAN among the member links.

In the following sections, concept of RBridge Group, together with its nickname, is introduced to fix these issues.

3. Concept of Virtual RBridge and Pseudo-nickname

A Virtual RBridge (RBv) represents a group of different ports on different edge RBridges, on which these RBridges provide end-station service to a set of their attached end stations. After joining RBv, such an RBridge port is called a member port of RBv, and such an RBridge becomes a member RBridge of RBv. In an RBridge RBv is identified by its virtual nickname in TRILL campus, and virtual nickname is also referred to as pseudo-nickname in this specification.

After joining an RBv, a member RBridge will announce its connection to RBv by including the information of that RBv, e.g., the pseudo-nickname of RBv, in its self-originated LSP. From such LSPs, remote RBridges that are not a members of RBv can deduce that, one or more shortest paths are available from itself to RBv.

When receiving a native frame on such a port, the member RBridge uses the RBv's nickname, instead of its own nickname, as ingress nickname in TRILL header if necessary to encapsulate the frame into TRILL data form. By de-capsulating such a TRILL-encapsulated data frame, a remote RBridge learns that S is reachable through RBv.

NOTE: An RBridge port can join at most one RBv at any time, but different ports on the same RBridge can join the same RBv or different RBvs. After joining an RBv, such a port becomes a member port of the RBv, and the RBridge becomes a member RBridge of the RBv. Furthermore, for a member RBridge, it MUST move out of RBv and clear the RBv's information from its self-originated LSPs when it loses the last member port from this group, due to port down, configuration, etc.

Use of the Appointed Forwarder framework specified in [RFC6325], this specification allows to utilize a single framework for both shared LAN and point-point edge connectivity. Additionally this allows to

- o Detect and protect against mis-configuration at the edge, e.g., on the device SW1 the two interfaces are not configured as LAG or

- o Accept TRILL and native frames on the RBridge interface connecting S1 in above Figure 1.

- o Avoid loops in the event S1 and S2 were connected by a native Ethernet Link.

3.1. VLAN-x Appointed Forwarder for member interfaces in RBv

If member RBridges in RBv cannot see each others' Hellos on their member ports (e.g., in the LAG scenario), then each RBridge becomes Designated RBridge (DRB) for that port and appoints itself as AF for all VLANs as it does not see any TRILL hellos on that port. However, it MAY acts as appointed forwarders only for parts of VLANs on that port, if it knows explicitly the sets of service VLANs on that port via other means. For example, administrator can statically configured the sets of service VLANs on that port, or a lower protocol (e.g., LAG protocol) informs TRILL the sets of service VLANs on that port, etc.

However, if they can see each others' Hellos on the member ports in RBv (e.g., in the shared link scenario), the TRILL Hello protocol in [RFC6325] is used for DRB election and for VLAN-x AFs appointment on those ports. Then the DRB appoints different member ports as AFs for different sets of VLANs.

Among the member RBridges of RBv, only the VLAN-x forwarder is responsible to ingress native traffic (both unicast and non-unicast traffic) in this VLAN into TRILL campus, but non-forwarder member RBridge is also permitted to egress unicast TRILL data traffic out of TRILL campus. For the multi-destination TRILL data frames, only the VLAN-x forwarder can egress their out of TRILL campus.

3.2. Announcing Pseudo-Nickname of RBv

Each member RBridge advertises the RBv's pseudo-nickname using the nickname sub-TLV [rfc6326bis], along with its regular nickname or nicknames, in its LSPs. When a member RBridge leaves from RBv due to losing its last member ports in RBv, it MUST clear RBv's pseudo-nickname from its update LSPs.

4. Distribution Trees for Member RBridges in RBv

In TRILL, RBridges use distribution trees to forward multi-destination frames. When a native frame either to destinations whose

location is unknown or to multicast/broadcast groups is necessary to be ingressed into TRILL campus, the ingress RBridge encapsulates it into multi-destination TRILL data frame and forwards it along a chosen distribution tree. In the TRILL header of the TRILL frame, the ingress nickname identifies the ingress RBridge and the egress nickname represents the root of the chosen tree. After receiving a multi-destination TRILL data frame, the RBridge performs Reverse Path Forwarding (RPF) check, along with other checks, on the multi-destination frame to further control potentially looping traffic.

RPF specifies that a multi-destination TRILL data frame ingressed by an RBridge and forwarded along a distribution tree can only be received by an RBridge on an expected port. If not on that port, that frame MUST be dropped by that RBridge.

However, member RBridges employ RBv's pseudo-nickname other than their own nicknames as ingress nickname when they ingress native frames received on member ports, regardless unicast or non-unicast frames, into TRILL campus. Therefore, when these frames reach a remote RBridge, they will be treated, by that RBridge, as frames ingressed by the same RBridge, i.e., RBv. If they are multi-destination frames and the same distribution tree is chosen by different member RBridges to forward these frames, they will travel along the tree and reach a remote RBridge on different ports. Then the RPF check is violated, and some of the frames reaching the RBridge on unexpected ports are dropped by the RBridge.

To fix the above issue, a scalable and resilient approach is proposed in [CMT], where different member RBridges are assigned different distribution trees for forwarding the multi-destination TRILL data frames that using RBv's pseudo-nickname as ingress nickname in their TRILL header. And a new TLV, named Affinity sub-TLV, is also introduced for a member RBridge to announce its assigned distribution tree for RBv in its self-originated LSPs. After receiving such LSPs, remote RBridges can calculate their RPF check information for RBv on those specified trees.

In this specification, the approach proposed in [CMT] is employed for RBv to assign different distribution trees to different member RBridges and the Affinity sub-TLV for member RBridges to announce their assigned trees in LSPs.

When a member RBridge joins into or leaves from a virtual RBridge group RBv due to its last member ports up/down or its configuration changing, etc., the distribution trees assigned to different member RBridges may change. That change and its influence on frame processing are beyond the scope of this document.

5. Frame Processing

Although, there are five types of Layer 2 frames in [RFC6325], e.g., native frame, TRILL data frame, TRILL control frames, etc., pseudo-nickname of RBv is only used for native frame and TRILL data frame in this specification.

5.1. Data Frames

5.1.1. Native Frames Ingressing

When RB1 receives a native frame on one of its valid member ports of RBv, it uses the pseudo-nickname of RBv, instead of its own nickname, as ingress nickname, if it is the appointed forwarder for the VLAN of the frame on the port. If the frame is not received on a member port, RB1 MUST NOT use RBv's pseudo-nickname as ingress nickname when doing TRILL-encapsulation on the frame. Otherwise, the reverse traffic may be forwarded to another member RBridge that does not connect to the link containing the destination, which may cause the traffic disruption.

If the above native frame is ingressed by RB1 as a multi-destination TRILL data frame, e.g., its destination is unknown to RB1 or it is non-unicast frame, RB1 can only choose one of its assigned distribution trees for RBv to distribute the TRILL-encapsulated frame [CMT]. If not so, the multi-destination TRILL data frame will fail RPF check on another RBridge and be dropped.

Furthermore, for such a frame, its source MAC address information ({ VLAN, Outer.MacSA, port }) is learned by default if its source address is unicast. Then the learned information is shared with other member RBridges of RBv (See Appendix A for more details for the information sharing).

5.1.2. TRILL Data Frames Egressing

This section describes egress processing of the received TRILL data frames on a member RBridge(RBn, say) in virtual RBridge group of RBv. Section 5.1.2.1 describes unicast TRILL data frames egress processing. Section 5.1.2.2 covers multi-destination TRILL data frames egress processing.

5.1.2.1. Unicast TRILL Data Frames

When receiving a unicast TRILL data frame, RBn checks the egress nickname in the TRILL header of the frame. If the egress nickname is one of RBn's own nicknames, the frame is processed as Section 4.6.2.4 in [RFC6325].

If the egress nickname is RBv's pseudo-nickname and RBn is a valid member RBridge of RBv, the Inner.MacSA and Inner.VLAN ID are, by default, learned associated with the ingress nickname, unless that nickname is unknown or reserved or the Inner.MacSA is not unicast. If the learned {Inner.MacSA, Inner.VLAN ID, ingress nickname} triplet is a new one or updates the locally stored one, this triplet is shared with other member RBridges of RBv (See Appendix A for more details for the triplet sharing).

Then the frame being forwarded is de-capsulated to native form. The Inner.MacDA and Inner.VLAN ID are looked up in RBn's local forwarding address cache, and one of the three following cases occurs:

- o If the destination end station identified by the Inner.MacDA and Inner.VLAN ID is on a local link to RBv, this frame is sent onto the link containing the destination.

- o else if RBn can reach the destination through another RBridge RBk, it re-encapsulate the native frame into a unicast TRILL data frame and sends it to RBk. RBn uses RBk's own nickname, instead of RBv's pseudo-nickname as egress nickname for the re-encapsulation, and remains the ingress nickname unchanged.

- o Else, RBn does not know how to reach the destination; it sends the native frame out of all its member ports of RBv on which it is appointed forwarders for the Inner.VLAN.

5.1.2.2. Multi-Destination TRILL Data Frames

If the RBn is an appointed forwarder for the Inner.VLAN ID of the frame, the Inner.MacSA and Inner.VLAN ID are, by default, learned as associated with the ingress nickname unless that nickname is unknown or reserved or the Inner.MacSA is not unicast. If the learned {Inner.MacSA, Inner.VLAN ID, ingress nickname} triplet is a new one or updates the locally stored one, this triplet is shared among the members RBridges in virtual RBridge group RBv (See Appendix A for more details for the triplet sharing).

Then a copy of the frame is de-capsulated into its native form. Before the native frame is sent out of ports on which RBn is appointed forwarder for the frame's VLAN, the following extra check is performed:

- o Assigned Distribution Trees Check (ADTC): If the flag for this check (ADTC_flag) is not zero on such a port, the distribution tree T along which the TRILL data frame arrives at RBn is checked. Only if T is one of RBn's assigned distribution trees in RBv, the native frame can be send out of this port. If not, the frame

cannot be sent out of this port.

The value of ADTC_flag on a RBridge's end-station-servicing port depends on whether the port is a member port of RBv and RBn can not receive Hellos from other member RBridges on that port or not.

If a port is a member port of RBv and RBn is the appointed forwarder for all VLANs on that port, the ADTC_flag MUST be set 1. For all other cases ADTC_flag MUST be set to zero.

6. Member Link Failure in RBv

In Figure 1, if the link SW1-RB1 fails, RB1 loses its only local link to S1. When that failure is detected, the MAC entries through the failed link to S1 are removed from RB1's forwarding table immediately, and other MAC entries to S1 shared by other member RBridges of RBv (See Appendix A for more details), e.g., RB2, are installed into RB1's forwarding table when RB1 still has at least one valid member port in RBv. Then when the TRILL-encapsulated traffic to S1 is delivered to RB1, it can be re-encapsulated by RB1 and forwarded, based on the available MAC entries, to another member RBridge which has direct link to S1 and egresses the traffic to S1.

On the other hand, if RB1 has lost all its member ports of RBv, it MUST updates its self-orignated LSPs to announce its giving up of membership of RBv and no longer utilizes pseudo-nickname of RBv to ingress/egress traffic into/out of TRILL campus until one of its member ports of RBv becomes valid.

NOTE: Although on an edge RBridge different ports that connect to different LAGs or LANs can join the same RBv, for simplicity, it is RECOMMENDED that on an edge RBridge different ports connecting to different LAGs or LANs join different RBvs in practical deployment, each RBv per LAG or per LAN. Then for such an edge RBridge, when all its member ports connecting to a LAG or LAN failed, it can move out of this RBv and no longer uses the RBv's pseudo-nickname to ingress/ egress data traffic into/out of TRILL campus.

7. OAM Frames

Special attention must be paid when generate the OAM frames. When an OAM frame is generated with ingress nickname of RBv, originator RBridge's nickname MUST be included in the OAM message to ensure response is returned to the originating member of RBv group.

8. Configuration Consistency

It is important that VLAN membership of member ports of end switch SW1 is consistent across all of the member ports it is attaching to member RBridges of RBv in the point-piont scenario. Any inconsistencies in VLAN membership may result in packet loss or having to through an extra hop RBridge before the packet reaches its destination end station.

As an example consider, in Figure 1, on RB1 link SW1-RB1 has VLAN1 and VLAN2 configured. Consider only VLAN1 is configured on RB2 on SW1-RB2 link. Both RB1 and RB2 use the same ingress nickname RBv for all frames originating from S1. Hence, RBx will learn MAC address from S1 on VLAN2 as originating from RBv. As a result, on the return path RBx may deliver VLAN2 traffic to RB2. RB2, does not have VLAN2 configured on SW1-RB2 link and hence may drop the frame or has to forward the frame to RB1 to egress it to S1 if RB2 knows RB1 can reach S1 in VLAN2.

9. IANA Considerations

TBD.

10. Security Considerations

TBD.

11. Acknowledgements

We would like to thank Mingjiang Chen for his contributions to this document. Additionally, we would like to thank Erik Nordmark, Les Ginsberg, Ayan Banerjee, Dinesh Dutt, Anoop Ghanwani, Janardhanan Pathang, and Jon Hudson for their good questions and comments.

12. Normative References

[CMT] Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-trill-cmt-00.txt Work in Progress, April 2012.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

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

[rfc6326bis] Eastlake 3rd, D., Banerjee, A., Ghanwani, A., and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-07.txt Work in Progress, March 2012.

Appendix A. Reasons for MAC Sharing among Member RBriges

With the introduction of virtual RBridge, MAC flip-flopping problem in LAN or LAG is resolved. However, in order to forward traffic effectively, member RBridges should share some of their learned MAC addresses with each other. For example, see Figure 2 shown below.

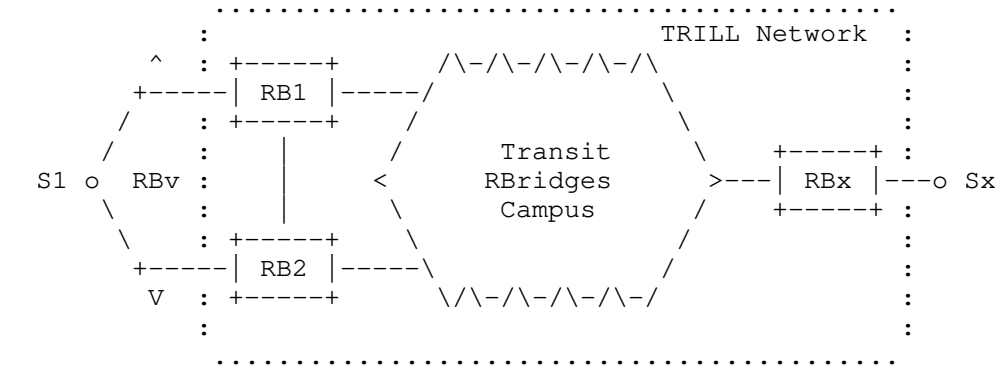


Figure 2 RBv in LAG scenario

In VLAN-x, native frames from S1 to Sx will enter TRILL campus through one member RBridge of the RBv, such as RB1 in Figure 2, so RB1 learns the location of S1 in VLAN-x; but with regard to reverse traffic, RBx may deliver it to RB2 if it thinks the shortest path to RBv is through RB2. Then, if RB2 knows the location of S1 and the link RB2-S1 is good, it egresses the traffic directly to S1. However, if the link fails and RB2 has not learn the location of S1 in VLAN-x, RB2 cannot transmit the traffic properly to S1.

Thus, the MAC addresses of attached end stations on one member RBridge SHOULD be shared with the rest member RBridges in an RBv. With these informations shared, when RB2 receives reverse frames, it can determine how to forward them to S1, for example, forward them to RB1 if the link RB2-S1 fails.

On the other hand, RBx always delivers the reverse traffic to RB2 if it thinks the shortest path to RBv is through RB2. Then RB2 egresses the traffic and learns the location of Sx in the case its link to S1 is good. RB1 will not know where Sx is if neighbor it has other ways to get the location of Sx nor RB2 shares this information with it. As a result, it has to always treat the traffic from S1 to Sx as unknown destination traffic and multicast it in TRILL. Always multi-casting such traffic adds additional forwarding burden on TRILL network.

Therefore, in addition to local attached end station MAC addresses, the learned remote MAC addresses should also be shared among all other member RBridges in an RBv. With such information sharing, RB1 can treat the traffic to Sx as known destination traffic and unicast it to RBx.

Although we can extend ESADI (End Stations Address Distribution Information) protocol or LAG protocol, etc., for such MAC sharing, ways for the sharing are beyond the scope of this document.

Authors' Addresses

Hongjun Zhai
ZTE
68 Zijinghua Road, Yuhuatai District
Nanjing, Jiangsu 210012
China

Phone: +86 25 52877345
Email: zhai.hongjun@zte.com.cn

Tissa Senevirathne
Cisco Systems
375 East Tasman Drive
San Jose, CA 95134
USA

Phone: +1-408-853-2291
Email: tsenevir@cisco.com

Radia Perlman
Intel Labs
2200 Mission College Blvd
Santa Clara, CA 95054-1549
USA

Phone: +1-408-765-8080
Email: Radia@alum.mit.edu

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

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

Mingui Zhang
Huawei
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: zhangmingui@huawei.com

Fangwei Hu
ZTE
889 Bibo Road, Pudong District
Shanghai, Shanghai 201203
China

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

TRILL working group
Internet Draft
Category: Informational

L. Dunbar
D. Eastlake
Huawei
Padia Perlman
Intel
Igor Gashinsky
Yahoo

Expires: April 2013

October 22, 2012

TRILL Edge Directory Assistance Framework

draft-ietf-trill-directory-framework-01

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/ietf/1id-abstracts.txt.

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

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this

document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

Edge RBridges currently learn the mapping between MAC addresses and their egress RBridges by observing the data packets traversed through. When an ingress RBridge receives a data frame with its destination address (MAC&VLAN) unknown, the data frame is flooded within the VLAN across the TRILL campus. When there are more than one RBridge ports connected to one bridged LAN, only one of them can be designated as the Appointed Forwarder port for forwarding/receiving native traffic to/from each VLAN, the other RBridge ports on that LAN have to be disabled for native traffic in that VLAN.

This draft describes the framework for using directory services to assist edge RBridges by reducing multi-destination frames, particularly unknown unicast frames flooding, and ARP/ND, improving TRILL network scalability in environment, such as data centers.

Conventions used in this document

The terms ''Subnet'' and ''VLAN'' are used interchangeably in this document because it is common to map one subnet to one VLAN. The terms ''TRILL switch'' and ''RBridge'' are used interchangeably in this document.

Table of Contents

1. Introduction 4
2. Terminology 5
3. RBridge Campus Impact of Massive Number of End Stations in a DC5
   3.1. Issues of Flooding Based Learning in DCs 5
   3.2. Some Examples 7
4. Benefits of Directory Assisted Edge RBridge in DC 8
5. Generic operation of Directory Assistance 9
   5.1. Information in Directory for Edge Bridges 9
   5.2. Push Model 10
   5.3. Pull Model 11
6. Conclusion and Recommendation..... 12
7. Security Considerations..... 12
8. IANA Considerations 12

9. Acknowledgements 13
10. References 13
10.1. Normative References..... 13
10.2. Informative References..... 13
Authors' Addresses 13

1. Introduction

Edge RBridges (devices implementing [RFC6325], also known as TRILL Switches) currently learn the mapping between destination MAC addresses and their egress RBridges by observing data packets. When ingress RBridge receives a data frame with its destination address (MAC&VLAN) unknown, the data frame is flooded within that VLAN across the TRILL campus. When there are more than one RBridge ports connected to one bridged LAN, only one of them can be designated as the Appointed Forwarder port for forwarding/receiving native traffic to/from for each VLAN. The other RBridge ports on that LAN have to be blocked for native traffic in that VLAN. (This ''blocked'' state has no effect of TRILL Data or IS-IS frames, which can still be sent and received. It only affects native frames.)

This draft describes the framework for using directory services to assist edge RBridges by reducing multi-destination frames, particularly ARP, ND, and unknown unicast, improving TRILL network scalability in environments, such as data centers.

Data center networks are different from enterprise campus networks in several ways, in particular:

1. Data centers, especially Internet and/or multi-tenant data centers tend to have a large number of end stations with a wide variety of applications.
2. Topology is usually based on racks and rows.
   -Guest OSs assignment to Servers, Racks, and Rows is orchestrated by a Server/VM Management system, not at random.
3. Rapid workload shifting in data centers can accelerate the frequency of the physical servers being re-loaded with different applications. Sometimes, the applications loaded to one physical server at different times can belong to different subnets.
4. With server virtualization, there is an ever-increasing trend to dynamically create or delete VMs when demand for resource changes, to move VMs from overloaded servers to less loaded servers, or to aggregate VMs onto fewer servers when demand is light.

Both 3) and 4) above can lead to applications in one subnet being placed in different locations (racks or rows) or one rack having applications belonging to different subnets.

This draft describes why and how Data Center TRILL networks can be optimized by utilizing a directory assisted approach.

2. Terminology

Bridge: IEEE Std 802.1Q-2011 compliant device [802.1Q]. In this draft, Bridge is used interchangeably with Layer 2 switch.

DA: Destination Address

DC: Data Center

EoR: End of Row switches in data center. Also known as Aggregation switches in some data centers

FDB: Filtering Database for Bridge or Layer 2 switch

End Station: Guest OS running on a physical server or on a virtual machine. An end station has at least one IP address and at least one MAC address, which could be in DA or SA field of a data frame.

RBridge: A device implementing the TRILL protocol [RFC6325]

RSTP: Rapid Spanning Tree Protocol

SA: Source Address

Station: A node, or a virtual node, with IP and/or MAC addresses, which could be in the DA or SA of a data frame.

STP: Spanning Tree Protocol

ToR: Top of Rack Switch in data center. It is also known as access switches in some data centers.

VM: Virtual Machines

3. RBridge Campus Impact of Massive Number of End Stations in a DC

3.1. Issues of Flooding Based Learning in DCs

It is common for Data Center networks to have multiple tiers of switches, for example, one or two Access Switches for each server rack (ToR), aggregation switches for some rows (or EoR switches), and some core switches to interconnect the aggregation switches. Many aggregation switches deployed in data centers have high port

density. It is not uncommon to see aggregation switches interconnecting hundreds of ToR switches.

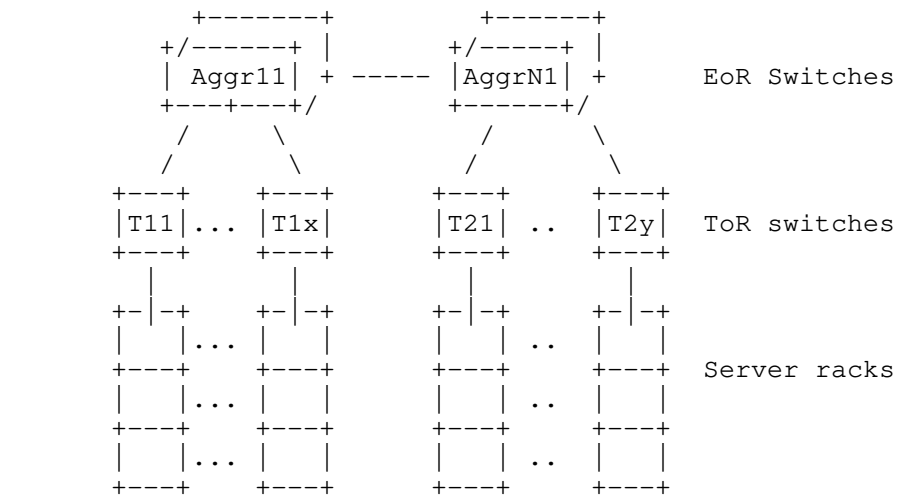


Figure 1: Typical Data Center Network Design

The following problems could occur when TRILL is deployed in a data center with large number of end stations, and the end stations in one subnet/VLAN could be placed under multiple edge RBridges:

- Unnecessary filling of slots in MAC table of edge RBridges RB1, due to RB1 receiving broadcast/multicast traffic (e.g. ARP/ND, cluster multicast, etc.) from end stations under other edge RBridges that are not actually communicating with any end stations attached to RB1.
- Some edge RBridge ports being blocked for user traffic when there are more than one RBridge ports connected to an edge bridged LAN. When there are multiple RBridge ports connected to a bridged LAN in each VLAN, only one (the Appointed Forwarder port) can forward/receive native traffic for that bridged LAN or VLAN. The rest have to be blocked for forwarding/receiving native traffic for that VLAN. When servers have dual uplinks to two different ToR switches (or edge RBridges), some links may not be fully utilized.
- Packets being flooded across TRILL campus when their DAs are not in ingress RBridge's cache.
- In an environment where VMs migrates, there is higher chance of cached entries becoming invalid, causing traffic to be black

holed by the egress RBridge. If VMs send out gratuitous ARP/ND or IEEE Std 802.1Qbg's VDP messages upon arriving at new locations, the ingress nodes might not have the MAC entries for the newly arrived VMs, causing more unknown flooding.

3.2. Some Examples

Consider a data center with 1600 server racks. Each server rack has at least one ToR switch. The ToR switches are further divided into 8 groups, with each group being connected by a set of aggregation switches. There could be 4 to 8 aggregation switches in each set to achieve load sharing for traffic to/from server racks. If TRILL is deployed in this data center environment, let's consider the following two scenarios for the TRILL campus boundary:

- Scenario #1: TRILL campus boundary starts at ToR switches:

  If each server rack has one uplink to one ToR, there are 1600 edge RBridges. If each rack has dual uplinks to two ToR switches, then there will be 3200 edge RBridges

  In this scenario, the TRILL domain will have more than 1600 (or 3200) + 8*4 (or 8*8) nodes, which is a large IS-IS domain. Even though a mesh IS-IS domain can scale up to thousands of nodes, it is challenging for aggregation switches to handle IS-IS link state advertisement among hundreds of parallel ports.

- Scenario #2: TRILL campus boundary starts at the aggregation switches:

  With the same assumption as before, the number of nodes in the TRILL campus will be less than 100, and aggregation switches don't have to handle IS-IS link state advisements among hundreds of parallel ports.

  But bridged LANs are formed under the aggregation switches in this scenario. With aggregation switches being the RBridge edge nodes, multiple RBridge edge ports could be connected to one bridged LAN. To avoid potential loops, TRILL requires only one of multiple RBridge edge ports connected to each VLAN being designated as Appointed Forwarder [RFC6439], and other ports being blocked for native frames in that VLAN.

  There is also the possibility of loops on the bridged LAN attached to RBridge edge ports unless STP/RSTP is running.

Running traditional Layer 2 STP/RSTP on the bridged LAN in this environment may be overkill because the topology among the ToR switches and aggregation switches is very simple and may not allow loops.

In addition, the number of MAC&VLAN<->Egress RBridge Mapping entries to be learned and managed by RBridge edge node can be very large. In the example above, each edge RBridge has 200 edge ports facing the ToR switches. If each ToR has 40 downstream ports facing servers and each server has 10 VMs, there could be 200*40*10 = 80000 end stations attached. If all those end stations belong to 1600 VLANs (i.e. 50 per VLAN) and each VLAN has 200 end stations, then under the worst-case scenario, the total number of MAC&VLAN entries to be learned by the edge RBridge can be 1600*200=320000, which is very large.

4. Benefits of Directory Assisted Edge RBridge in DC

In data center environment, applications placement to servers, racks, and rows is orchestrated by Server (or VM) Management System(s). That is, there is a database or multiple databases (distributed model) that have the knowledge of where each application is placed. If the application location information can be fed to RBridge edge nodes, in some form of Directory Service, then RBridge edge nodes won't need to flood data frames with unknown DA across the TRILL campus.

Avoiding unknown unicast DA flooding to TRILL campus is especially valuable in data center environment because there is higher chance of an edge RBridge receiving packets with unknown unicast DA and broadcast/multicast messages due to VM migration and servers being loaded with different applications. When a VM is moved to a new location or a server is loaded with a new application with different IP/MAC addresses, it is more likely that the DA of data packets sent out from those VMs are unknown to their attached edge RBridges. In addition, gratuitous ARP (IPv4) or Unsolicited Neighbor Advertisement (IPv6) sent out from those newly migrated or activated VMs have to be flooded to other edge RBridges that have VMs in the same subnets.

The benefits of using directory assistance include:

- Avoid flooding unknown unicast DA across TRILL campus. The Directory enforced MAC&VLAN <-> Egress RBridge mapping table can determine if a data packet needs to be forwarded across TRILL campus.

When multiple RBridge edge ports are connected via a bridged LAN to end stations (servers/VMs), a directory assisted edge RBridge won't need to flood unknown unicast DA data frames to all ports of the edge RBridges in the frame's VLAN when it ingresses a frame. It can trust the directory to tell it where the DA is or it will discard the frame if the directory says the DA does not exist in the campus.

- Reduce flooding of decapsulated Ethernet frames with unknown MAC-DA to a bridged LAN connected to RBridge edge ports.

  When an RBridge receives a TRILL frame whose destination Nickname matches with its own, the normal procedure is for the RBridge to decapsulate the TRILL header and forward the decapsulated Ethernet frame to the directly attached bridged LAN. If the destination MAC is unknown, the normal Ethernet switch's flooding will occurs to the decapsulated Ethernet frame. With directory assistance, the egress RBridge can determine if DA in a frame matches with any end stations attached via the bridged LAN. Frames can be discarded if their DAs do not match.

- Reduce the amount of MAC&VLAN <-> Egress RBridge mapping maintained by edge RBridges. There is no need for an edge RBridge to keep MAC entries of remote end stations that don't communicate with the end stations locally attached.

5. Generic operation of Directory Assistance

5.1. Information in Directory for Edge Bridges

To achieve the benefits of directory service for TRILL, the corresponding directory server entries will need, at a minimum, the following logical attributes:

[IP, MAC, attached RBridge nickname, {list of interested RBridges}]

The {list of interested RBridges} would get populated when an RBridge queries for information, or pushed down from management systems. The list is used to notify those RBridges whose connectivity to VMs changes due to VM migration or link failures.

There can be two different models for RBridge edge node to be assisted by Directory Service: Push Model and Pull Model.

5.2. Push Model

Under this model, Directory Server(s) push down the MAC&VLAN <-> Egress RBridge mapping for all the end stations that might communicate with end stations attached to an RBridge edge node. Under this model, it is recommended that the ingress RBridge simply drops a data packet (instead of flooding to TRILL campus) if the packet's destination address can't be found in the MAC&VLAN<->Egress RBridge mapping table.

It may not be necessary for every edge RBridge to get the entire mapping table for all the end stations in a data center. There are many ways to narrow the full set down to a smaller set of remote end stations that communicate with end stations attached to an edge RBridge. A simple approach of only pushing down the mapping for the VLANs that have active end stations under an edge RBridge can reduce the number of mapping entries being pushed down.

However, the Push Model usually will push down more entries of MAC&VLAN<->Egress RBridge mapping to edge RBridges than needed. Under the normal process of edge RBridge cache aging and unknown DA flooding, rarely used mapping entries would have been removed. But it can be difficult for Directory Servers to predict the communication patterns among applications within one VLAN. Therefore, it is likely that the Directory Servers will push down all the MAC&VLAN entries if there are end stations in the VLAN being attached to the edge RBridge. This is a major disadvantage of the Push Model compared with the Pull Model described below.

In the Push Model, it is necessary to have a way for an RBridge node to request directory server(s) to start pushing down the mapping entries. This method should at least include the VLANs enabled on the RBridge, so that directory server doesn't need to push down the entire mapping entries for all the end stations in the data center. An RBridge node must be able to get mapping entries when it is initialized or restarted.

The detailed method and hand-shake mechanism between RBridge and Directory Server(s) is beyond the scope of this framework draft.

When directory server needs to push down a very large number of entries to edge RBridges, summarization should be considered. For example, with one edge RBridge Nickname being associated with all attached end stations' MAC addresses and VLANs as shown below:

Nickname1	VID-1	MAC1, MAC2, ,MACn
	VID-2	MAC1, MAC2, ,MACn
	...	MAC1, MAC2, ,MACn
Nickname2	VID-1	MAC1, MAC2, ,MACn
	VID-2	MAC1, MAC2, ,MACn
	....	MAC1, MAC2, ,MACn
-------	------	MAC1, MAC2, ,MACn
	...	MAC1, MAC2, ,MACn

Table 1: Summarized table pushed down from directory

Whenever there is any change in MAC&VLAN <-> Egress RBridge mapping, that can be triggered by end stations being added, moved, or de-commissioned, an incremental update can be sent to the edge RBridges which are impacted by the change. Therefore, something like a sequence number has to be maintained by directory servers and RBridges. Detailed mechanisms will be specified in a separate draft.

5.3. Pull Model

Under this model, an RBridge pulls the MAC&VLAN<->Egress RBridge mapping entry from the directory server when its cache doesn't have the entry. There are several options to trigger the pulling process. For example, the RBridge edge node can send a pull request whenever it receives an unknown DA, or the RBridge edge node can simply intercept all ARP/ND requests and forward them to the Directory Server(s) that has the information on where the target stations are located. The ingress RBridge can cache the mapping pulled down from the directory.

One advantage of the Pull Model is that edge RBridge can age out MAC&VLAN entries if they haven't been used for a certain configured period of time or a period of time provided by the Directory. Therefore, each edge RBridge will only keep the entries which are frequently used, so mapping table size can be smaller. Edge RBridges would query the Directory Server(s) for unknown DAs in data frames or ARP/ND and cache the response. When end stations attached to remote edge RBridges rarely communicate with the locally attached end stations, the corresponding MAC&VLAN entries would be aged out from the RBridge's cache.

RBridge waiting for response from Directory Servers upon receiving a data frame with unknown DA is similar to a L2/L3 boundary router waiting for ARP/ND response upon receiving an IP data frame whose DA is not in the router's IP/MAC cache table. Most deployed routers today do hold the packets and send an ARP/ND requests to the target upon receiving a packet with DA not in its IP-MAC cache. When ARP/ND replies are received, the router will send the data frame to the target. This practice is to minimize flooding when targets don't exist in the subnet.

When the target doesn't exist in the subnet, routers generally re-send ARP/ND request a few more times before dropping the packets. Therefore, the holding time by routers to wait for ARP/ND response can be longer than the time taken by the Pull Model to get IP-MAC mapping from directory if target doesn't exist in the subnet.

A separate draft will specify the detailed messages and mechanism for edge RBridge to pull information from directory server(s).

6. Conclusion and Recommendation

The traditional RBridge learning approach of observing data plane can no longer keep pace with the ever growing number of end stations in Data centers.

Therefore, we suggest TRILL consider directory assisted approach(es). This draft only describes the basic framework of using directory assisted approach for RBridge edge nodes. More complete mechanisms will be described in separate drafts.

7. Security Considerations

Accurate mapping of IP addresses into MAC addresses is important to the correct delivery of information. The security of specific directory assisted mechanisms will be discussed in the draft or drafts specifying those mechanisms.

For general TRILL security considerations, see [RFC6325].

8. IANA Considerations

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

9. Acknowledgements

This document was prepared using 2-Word-v2.0.template.dot.

10. References

10.1. Normative References

[RFC6352] Perlman, et, al ''RBridge: Base Protocol Specification'', https://datatracker.ietf.org/doc/rfc6325/, July, 2011

[RFC6439] Perlman, et, al ''RBridges: Appointed Forwarders'', https://datatracker.ietf.org/doc/rfc6439/, Nov 2011

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

10.2. Informative References

[802.1Q] IEEE Std 802.1Q-2011, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", May 2011.

[802.1Qbg] IEEE Std 802.1Qbg-2012, ''Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks -- - Edge Virtual Bridging'', July 2012.

Authors' Addresses

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA
Phone: 1-508-333-2270
Email: d3e3e3@gmail.com

Radia Perlman
Intel Labs
2200 Mission College Blvd.
Santa Clara, CA 95054-1549 USA
Phone: +1-408-765-8080
Email: Radia@alum.mit.edu

Igor Gashinsky
Yahoo
45 West 18th Street 6th floor
New York, NY 10011
Email: igor@yahoo-inc.com

TRILL Working Group
INTERNET-DRAFT
Intended status: Proposed Standard
Updates: 6325, 6327

Donald Eastlake
Mingui Zhang
Huawei
Puneet Agarwal
Broadcom
Radia Perlman
Intel Labs
Dinesh Dutt
Cumulus Networks

Expires: April 20, 2012
October 21, 2012

TRILL: Fine-Grained Labeling
<draft-ietf-trill-fine-labeling-02.txt>

Abstract

The IETF has standardized TRILL (TRansparent Interconnection of Lots of Links), a protocol for least cost transparent frame routing in multi-hop networks with arbitrary topologies and link technologies, using link-state routing and encapsulation with a hop count.

The TRILL base protocol standard supports labeling of TRILL data with up to 4K IDs. However, there are applications that require more fine-grained labeling of data. This document updates RFC 6325 and 6327 by specifying extensions to the TRILL base protocol to safely accomplish this.

Status of This Memo

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

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list.

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/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.....3
1.1 Terminology.....3
1.2 Contributors.....4

2. Fine-Grained Labeling.....5
2.1 Goals.....5
2.2 Base Protocol TRILL Data Labeling.....6
2.3 Fine-Grained Labeling (FGL).....7

3. Campus Wide VL versus FGL Semantic Differences.....9
4. Interaction with VL TRILL Switches.....10

5. Fine-Grained Labeling Details.....12
5.1 Ingress Processing.....12
5.2 Transit Processing.....12
5.2.1 Unicast Transit Processing.....13
5.2.2 Multi-Destination Transit Processing.....13
5.3 Egress Processing.....13
5.4 Appointed Forwarders and the DRB.....14
5.5 Address Learning.....14
5.6 ESADI Extensions.....14

6. IS-IS Extensions.....16
7. Comparison to Goals.....17

8. Allocation Considerations.....18
8.1 IEEE Allocation Considerations.....18
8.2 IANA Considerations.....18

9. Security Considerations.....19

Acknowledgements.....19
Normative References.....20
Informative References.....20
Change History.....21

1. Introduction

The IETF has standardized the TRILL (TRansparent Interconnection of Lots of Links) protocol [RFC6325]. TRILL switches provide a solution for least cost transparent routing in multi-hop networks with arbitrary topologies and link technologies, using [IS-IS] [RFC6165] [RFC6326bis] link-state routing and encapsulation with a hop count. They address the problems outlined in [RFC5556]. TRILL switches are sometimes called RBridges (Routing Bridges).

The TRILL base protocol standard supports labeling of TRILL data with up to 4K IDs. However, there are applications that require more fine-grained labeling of data for configurable isolation based on different tenants, service instances, or the like. This document updates [RFC6325] and [RFC6327] by specifying extensions to the TRILL base protocol to safely accomplish this.

Familiarity with [RFC6325] and [RFC6326bis] is assumed in this document.

1.1 Terminology

The terminology and acronyms of [RFC6325] are used in this document with the additions listed below.

DEI - Drop Eligibility Indicator [802.1Q]

FGL - Fine-Grained Labeling or Fine-Grained Labeled or Fine-Grained Label

FGL RBridge - A TRILL switch that support both FGL and VL

Edge RBridge - A TRILL switch announcing VL or FGL connectivity in its LSP

TRILL Switch - Alternative name for an RBridge

VL - VLAN Labeling or VLAN Labeled or VLAN Label

VL RBridge - A TRILL switch that supports VL but does not support FGL

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].

1.2 Contributors

Thanks for the contributions of the following:

Tissa Senevirathne

2. Fine-Grained Labeling

The essence of Fine-Grained Labeling (FGL) is that (a) when TRILL Data frames are ingressed or created they may incorporate a label from a set consisting of significantly more than 4K labels, (b) TRILL switch (RBridge) ports can be labeled with a set of such labels, and (c) an FGL TRILL Data frame cannot be egressed through an RBridge port unless its fine-grained label (FGL) matches one of the labels of the port.

Section 2.1 lists FGL goals. Section 2.2 briefly outlines the more coarse TRILL base protocol standard [RFC6325] data labeling. And Section 2.3 outlines a method of FGL of TRILL Data frames.

2.1 Goals

There are several goals that should be met by FGL in TRILL. They are briefly described in the list below in approximate order by priority with the most important first.

1. Fine-Grained

   Some networks have a large number of entities that need configurable isolation, whether those entities are independent customers, applications, or branches of a single endeavor or some combination of these or other entities. The labeling supported by [RFC6325] provides for only (2**12 - 2) valid identifiers or labels. A substantially larger number is required.

2. Silicon Considerations

   Fine-grained labeling (FGL) should, to the extent practical, use existing features, processing, and fields that are already supported in at least some fast path silicon implementations that currently support the TRILL base protocol.

3. Base RBridge Compatibility

   To support some incremental conversion scenarios, it is desirable that not all RBridges in a campus using FGL be required to be FGL aware. That is, it is desirable that RBridges not implementing the FGL feature and performing at least the transit forwarding function can usefully process TRILL Data frames that incorporate FGL.

4. Alternate Priority

It would be desirable for an ingress TRILL Switch to be able to assign a different priority to an FGL TRILL Data frame for its ingress-to-egress propagation from the priority of the original native frame. The original priority should be restored on egress.

2.2 Base Protocol TRILL Data Labeling

This section provides a brief review of the [RFC6325] TRILL Data frame internal VL Labeling and changes the description of the TRILL Header by moving its end point. This description change does not involve any change in the bits on the wire or in the behavior of existing [RFC6325] RBridges.

Currently TRILL Data frames have the VL structure shown below:

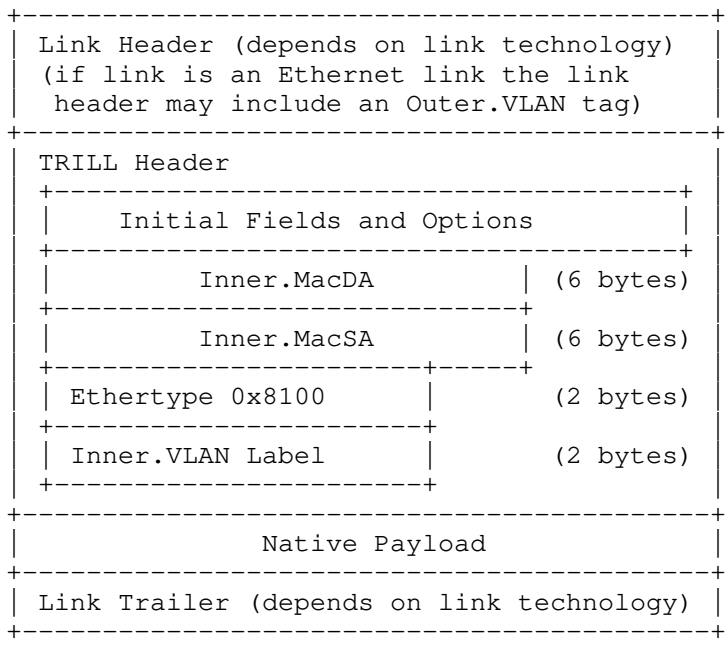


Figure 1. TRILL Data with VL

In the base protocol as specified in [RFC6325] the 0x8100 value is always present and is followed by the Inner.VLAN field which includes the 12-bit VL.

2.3 Fine-Grained Labeling (FGL)

FGL expands the data label available under the TRILL base protocol standard to a fine-grained label (FGL) with a 12-bit high order part and a 12-bit low order part. In this document, FGLs are usually denoted as "(X.Y)" where X is the high order part and Y is the low order part of the FGL. The FGL information appears in the TRILL Header as shown below.

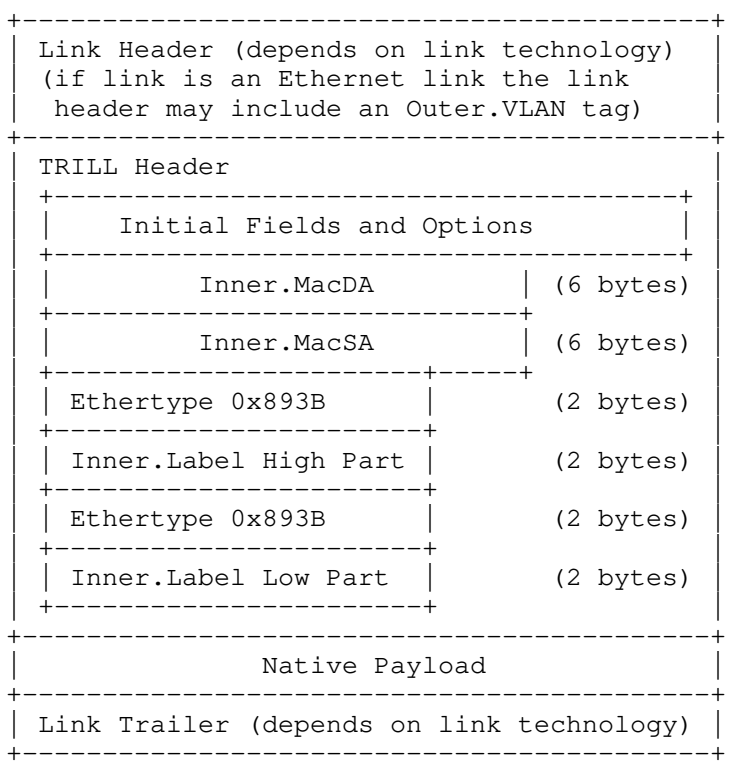


Figure 2. TRILL Data with FGL

For FGL frames, the inner MAC address fields are followed by the FGL information using 0x893B.

The two bytes following each 0x893B have, in their low order 12 bits, fine-grained label information. The upper 4 bits of those two bytes are used for a 3-bit priority field and one drop eligibility indicator (DEI) bit.

The priority field of the Inner.Label High Part is the priority used for frame transport across the TRILL campus from ingress to egress. The label bits in the Inner.Label High Part are the high order part of the FGL and those bits in the Inner.Label Low Part are the low

order part of the FGL.

The appropriate FGL value for an ingressed native frame is determined by the ingress RBridge port as specified in Section 5.1. Ports of TRILL switches supporting FGL also have capabilities to transmit frames being forwarded or egressed as untagged or VL as specified in Section 5.3.

3. Campus Wide VL versus FGL Semantic Differences

There are differences between the semantics across a TRILL campus for VL and FGL labeled TRILL Data frames.

With VL, data label IDs have the same meaning throughout the campus and are from the same label space as the VLAN IDs used on Ethernet links to end stations.

With TRILL FGL, many things remain the same. Ports of FGL TRILL switches, up through the usual VLAN and priority processing, act as they do for VL TRILL switches: Ethernet links between FGL TRILL switches still have only C-VLAN tagging on them and TRILL switch ports provide a VLAN ID for an incoming frame and accepts a VLAN ID for a frame being queued for output. Appointed Forwarders [RFC6439] on a link are still appointed for a C-VLAN. The Designated VLAN for an Ethernet link is still a C-VLAN.

The larger FGL space is a different space from the VL data label space. For ports configured for FGL, the C-VLAN on an ingressed native frame is mapped to the FGL data label space with a potentially different mapping for each port. A similar FGL to C-VLAN mapping occurs per port on egress. Thus, for ports configured for FGL, the native frame C-VLAN on one link corresponding to an FGL can be different from the native frame C-VLAN corresponding to that same FGL on a different link elsewhere in the campus or even a different link attached to the same RBridge. The FGL label space is flat and does not hierarchically encode any particular number of native frame C-VLAN bits or the like. FGLs in TRILL Data frames appear only inside the TRILL Header after the inner MAC addresses. They are only seen by TRILL aware devices.

FGL RBridge ports can be configured for FGL or VL with VL being the default. As with a base protocol [RFC6325] RBridge, an unconfigured FGL TRILL switch port reports an untagged frame it receives as being in VLAN 1.

4. Interaction with VL TRILL Switches

It is not possible for VL TRILL switches to handle FGL frames even if the VL TRILL switch is only acting in the transit capacity. This is because VL frames are required to have 0x8100 at the beginning of the data label where FGL TRILL switches have 0x893B. VL-only TRILL switches conformant to [RFC6325] should discard frames with this new value after the inner MAC addresses and, if they do not discard such frames, they will be confused (see Section 9 below).

If there are FGL TRILL switches in a campus, it is assumed that the intent is for all TRILL switches in that campus to support FGL. Any VL TRILL switches present are isolated by FGL TRILL switches as follows: FGL RBridges will report their FGL capability in LSPs. Thus FGL TRILL switches (and any management system with access to the link state database) will be able to detect the existence of TRILL switches in the campus that do not support FGL. If any such VL TRILL switches are present on a link then, although all other aspects of the adjacency machinery work as normal [RFC6327], any FGL TRILL switches on the link will not create a pseudo node for the link if they are DRB and do not announce any adjacencies they have on the link. As a result, although adjacencies between two or more VL RBridge ports on the link could become part of the campus topology and pass TRILL Data frames, no adjacency from an FGL RBridge port to a VL RBridge port or to a pseudonode will be reported for such a mixed FGL/VL link. Since an adjacency must be reported up by both ends before it becomes part of the campus topology, even though adjacencies to an FGL RBridge might be reported by a VL RBridge, no TRILL Data can flow between an FGL RBridge port and a VL RBridge port.

The usual DRB election operates on a link with mixed FLG and VL ports. If an FGL RBridge port is DRB, it MUST handle all native traffic or appoint only other FGL ports as forwarder for one or more VLANs, so that all end stations will get service to the FGL campus. If a VL RBridge port is DRB, it will not understand that FGL RBridge ports are different. To the extent that a VL DRB handles native frames or appoints other VL ports on a link to handle native frames for one or more VLANs, the end stations sending and receiving those native frames will be isolated from the FGL campus. To the extent that a VL DRB happens to appoint an FGL port as Appointed Forwarder for one or more VLANs, the end stations sending and receiving native frames in those VLANs will get service to the FGL campus. This somewhat odd corner case behavior is considered acceptable because it is assumed that VL TRILL switches in an FGL campus are infrequent misconfigurations.

For links configured as point-to-point, if the TRILL switches at each end are both VL or both FGL, a bi-directional adjacency can be formed by the usual mechanisms. If one is VL and one is FGL but the point-

to-point link is otherwise correctly configured, the VL TRILL switch will report an adjacency but the LFG one will not. As a result, the link will not become part of the topology and TRILL Data cannot flow over the link, isolating the VL TRILL switch.

5. Fine-Grained Labeling Details

This section specifies ingress, transit, egress, and other processing of TRILL Data frames with regard to FGLs. A transit or egress FGL TRILL switch determines that a TRILL Data frame is FGL by detecting that the inner MAC address fields are followed by 0x893B.

5.1 Ingress Processing

An FGL RBridge MAY be configured, on one or more ports, to ingress native frames as FGL. Any ports not so configured that accepts a native frame perform the previously specified VL ingress processing on native frames [RFC6325]. There is no change in Appointed Forwarder logic (see Section 5.4).

FGL TRILL switches MUST support configurable per port mapping from the VL of a native frame, as reported by the ingress port, to an FGL. FGL TRILL switches MAY support other methods to determine the FGL of an incoming native frame, such as based on the protocol of the native frame or local knowledge.

The FGL ingress process MUST place the priority and DEI associated with an ingressed native frame in upper 4 bits of the Inner.Label Low Order part. It SHOULD also associate a possibly different mapped priority and DEI with an ingressed frame. The mapped priority is placed in the Inner.Label High Part. If such mapping is not supported then the original priority and DEI MUST be placed in the Inner.Label High Part.

An FGL ingress RBridge MAY serially TRILL unicast a multi-destination TRILL Data frame to the relevant egress TRILL switches after encapsulating it as a TRILL known unicast data frame (M=0) and SHOULD unicast such a multi-destination TRILL Data frame if there is only one relevant egress FGL RBridge. For FGL RBridges, this permits serial unicast of multi-destination frames by the ingress as an alternative to the use of a distribution tree. The relevant egress TRILL switches are determined by starting with those announcing connectivity to the frame's (X.Y) label. That set SHOULD be further filtered based on multicast listener and multicast router connectivity if the native frame was a multicast frame.

5.2 Transit Processing

TRILL Data frame transit processing is fairly straightforward as described in Section 5.2.1 for known unicast TRILL Data frames and in Section 5.2.2 for multi-destination TRILL Data frames.

5.2.1 Unicast Transit Processing

There is very little change in TRILL Data frame unicast transit processing. A transit TRILL switch forwards any unicast TRILL Data frame to the next hop towards the egress RBridge as specified in the TRILL Header. All transit TRILL switches, whether VL or FGL, MUST take the priority and DEI used to forward a frame from the Inner.VLAN label or the FGL Inner.Label High Part. These bits are in the same place in the frame.

5.2.2 Multi-Destination Transit Processing

Multi-destination TRILL Data frames are forwarded on a distribution tree selected by the ingress TRILL switch except that an FGL ingress RBridge MAY choose to TRILL unicast such a frame to all relevant egress TRILL switches. The distribution trees do not distinguish between FGL and VL multi-destination frames except, possibly, in pruning behavior. All distribution trees are calculated as provided for in the TRILL base protocol standard [RFC6325]. There is no change in the Reverse Path Forwarding Check.

An FGL RBridge, say RB1, having an FGL multi-destination frame for label (X.Y) to forward on a distribution tree, SHOULD prune that tree based on whether there are any edge TRILL switches on a tree branch that are advertising connectivity to label (X.Y). In addition, RB1 SHOULD prune multicast frames based on reported multicast listener and multicast router attachment in (X.Y).

Pruning is an optimization. If a transit TRILL switch does less pruning than it could, there may be greater link utilization than strictly necessary but the campus will still operate correctly. For example, a transit TRILL switch could prune based on only part of the FGL such as only the High Part or only the Low Part.

5.3 Egress Processing

Egress processing is generally the reverse of ingress progressing described in Section 5.1.

An FGL RBridge MUST be able to configurably convert the FGL in an FGL TRILL Data frame it is egressing to a VLAN ID for the resulting native frame on a per port basis. A port MAY be configured to strip output VLAN tagging. It is the responsibility of the network manager to properly configure the TRILL switches in the campus to obtain the desired mappings.

The priority and DEI of the egressed native frame are taken from the Inner.Label Low Order Part.

An FGL RBridge egresses FGL frames similarly to the egressing of VL frames, as follows:

1. A known unicast FGL frame is egressed to the FGL port matching its fine-grained label and Inner.MacDA. If there is no such port, it is flooded out all FGL ports that have its FGL unless the TRILL switch has knowledge that the frame's Inner.MacDA cannot be out that port.

2. A multi-destination FGL frame is decapsulated and flooded out all ports with its FGL, subject to multicast pruning.

FGL RBridges MUST accept multi-destination encapsulated frames that are sent to them as TRILL unicast frames, that is, frames that may have a multicast or broadcast Inner.MacDA (or are being sent to an unknown unicast Inner.MacDA) and the TRILL Header M bit = 0. They locally egress such frames, if appropriate, but MUST NOT forward them (other than egressing them as native frames on their local links).

5.4 Appointed Forwarders and the DRB

There is no change in Adjacency [RFC6327] or Appointed Forwarder logic [RFC6439] on a link regardless of whether some or all the ports on the link are for FGL RBridges except as described in Section 4 above.

5.5 Address Learning

An FGL TRILL switch learns addresses on FGL ports based on the fine-grained label rather than the native frame's VLAN. Addresses learned from ingressed native frames on FGL ports are logically represented by { MAC address, fine-grained label, port, confidence, timer } while remote addresses learned from egressing FGL frames are logically represented by { MAC address, fine-grained label, remote TRILL switch nickname, confidence, timer }.

5.6 ESADI Extensions

The TRILL ESADI (End Station Address Distribution Information) protocol is specified in [RFC6325] as optionally transmitting MAC address connection information through TRILL Data frames between

participating TRILL switches over the virtual link provided by the TRILL multicast frame distribution mechanism. In [RFC6325], the VLAN to which an ESADI frame applies is indicated only by the Inner.VLAN label and no indication of that VLAN is allowed within the ESADI payload.

ESADI is extended to support FGL by providing for the indication of the FGL to which an ESADI frame applies only in the Inner.Label of that frame and no indication of that FGL is allowed within the ESADI payload.

6. IS-IS Extensions

Extensions to the TRILL use of IS-IS are required to support FGL include the following:

1. An method for a TRILL switch to announce itself in its LSP as supporting FGL.

2. A sub-TLV analogous to Interested VLANs and Spanning Tree Roots sub-TLV of the Router Capabilities TLV but indicating FGLs rather than VLs (see Section 8.2). This is called the Interested Labels and Spanning Tree Roots sub-TLV in [rfc6326bis].

3. Sub-TLVs analogous to the GMAC-ADDR sub-TLV of the Group Address TLV that specifies an FGL rather than a VL (see Section 8.2.). This are called the GLMAC-ADDR, GLIP-ADDR, and GLIP6 ADDR sub-TLVs in [rfc6326bis].

7. Comparison to Goals

Comparing TRILL FGL, as specified in this document, with the goals given in Section 2.1, we find as follows:

1. Fine-Grained: FGL provides 2**24 labels, vastly more labels than the 4K VL labels provided in [RFC6325].

2. Silicon Considerations: Existing TRILL fast path silicon chips can perform base TRILL Header insertion and removal to support ingress and egress. In addition, it is believed that most such silicon chips can also perform the native frame to FGL mapping and the encoding of the FGL as specified herein, as well as the inverse decoding and mapping. Some existing silicon can perform only one of these operations on a frame in the fast path and is thus not suitable to implement fast path TRILL FGL processing; however, other existing chips are believed to be able to perform both operations on the same frame in the fast path and are suitable for FGL implementation.

3. Base RBridge Compatibility: As described in Section 3, FGL is not compatible with TRILL switches conformant to the base specification RBridges [RFC6325].

4. Alternate Priority: The encoding specified in Section 2.3 provides for a new priority and DEI in the Inner.Label First Part and a place to preserve the original user priority and DEI in the Second Part, so it can be restored on egress.

8. Allocation Considerations

Allocations by the IEEE Registration Authority and IANA are listed below.

8.1 IEEE Allocation Considerations

The IEEE Registration Authority has assigned Ethertype 0x893B for use as the FGL Ethertype.

8.2 IANA Considerations

IANA is requested to allocate capability bit TBD in the TRILL-VER sub-TLV capability bits [RFC6326bis] to indicate an RBridge is FGL-capable.

9. Security Considerations

See [RFC6325] for general RBridge Security Considerations.

As with any communications system, end-to-end encryption and authentication should be considered for sensitive data.

Confusion between a frame with VL X and FGL (X.Y) is a potential problem if a VL RBridge did not check for the occurrence of 0x8100 (see Sections 2.2 and 2.3) and discard such a frame. Possible problems with such a VL RBridge would be as follows:

1. If it received a TRILL Data frame with FGL (X.Y) it could egress it to an end station in VLAN-X. The payload of such an egressed frame would appear to begin with Ethertype 0x893B which would likely be discarded by an end station. Nevertheless, such an egress would almost certainly be a violation of security policy.

2. If it received a multi-destination TRILL Data frame with FGL (X.Y) and it pruned the distribution tree, it would incorrectly prune it on the basis of VLAN-X. This could lead to the multi-destination data frame not being delivered to all of its intended recipients.

These two potential problems would only occur in the case of the misconfiguration of attaching such a VL RBridge to an FGL campus; however, there is protection against this in that FGL RBridges will not announce adjacency to VL RBridges (see Section 4). As a result, no TRILL data frames can be exchanged between VL and FGL RBridges and VL RBridges will be isolated for data puposes.

Acknowledgements

The comments and suggestions of the following are gratefully acknowledged:

Anoop Ghanwani, Sujay Gupta, Weiguo Hao, Jon Hudson, Yizhou Li, Vishwas Manral, Erik Nordmark, and Ilya Varlashkin.

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

Normative References

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

[802.1Q] - IEEE 802.1, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, May 2011.

[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.

[RFC6326bis] - Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-ietf-isis-rfc6326bis, work in progress.

Informative References

[RFC5556] - Touch, J. and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement", RFC 5556, May 2009.

[RFC6165] - Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.

[RFC6327] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Dutt, D., and V. Manral, "Routing Bridges (RBridges): Adjacency", RFC 6327, July 2011

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

Change History

From -00 to -01:

Update author info and make editorial changes.

From -01 to -02

1. Change the value after the inner MAC addresses for FGL frames from 0x8100 to 0x893B

2. As a consequence of item 1 above, for safety prohibit use for TRILL Data of links between FGL and VL RBridges, isolating any VL RBridges. Make appropriate changes throughout document, including Security Considerations section, based on this change.

3. Reference and contributor updates.

4. Various editorial changes.

Authors' Addresses

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

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

Mingui Zhang
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Puneet Agarwal
Broadcom Corporation
3151 Zanker Road
San Jose, CA 95134 USA

Phone: +1-949-926-5000
Email: pagarwal@broadcom.com

Radia Perlman
Intel Labs
2200 Mission College Blvd.
Santa Clara, CA 95054 USA

Phone: +1-408-765-8080
Email: Radia@alum.mit.edu

Dinesh G. Dutt
Cumulus Networks
1089 West Evelyn Avenue
Sunnyvale, CA 94086 USA

Email: ddutt.ietf@hobbesdutt.com

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

TRILL Working Group
Internet Draft
Intended status: Informational

Tissa Senevirathne
CISCO
David Bond
IBM
Sam Aldrin
Yizhou Li
Huawei
Rohit Watve
CISCO

October 20, 2012

Expires: April 2013

Requirements for Operations, Administration and Maintenance (OAM) in TRILL

draft-ietf-trill-oam-req-02.txt

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), 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/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on April 20,2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

OAM (Operations, Administration and Maintenance) is a general term used to identify functions and toolsets to troubleshoot and monitor networks. This document presents, OAM Requirements applicable to TRILL.

Table of Contents

1. Introduction....3
   1.1. Scope....3
2. Conventions used in this document....3
3. Terminology....3
4. OAM Requirements....4
   4.1. Data Plane....4
   4.2. Connectivity Verification....5
      4.2.1. Unicast....5
      4.2.2. Multicast....5
   4.3. Continuity Check....5
   4.4. Path Tracing....6
   4.5. General Requirements....6
   4.6. Performance Monitoring....7
      4.6.1. Packet Loss....7
      4.6.2. Packet Delay....8
   4.7. ECMP Utilization....8
   4.8. Security and Operational considerations....8
   4.9. Fault Indications....9
   4.10. Defect Indications....9
   4.11. Live Traffic monitoring....9
5. Security Considerations....10
6. IANA Considerations....10
7. References....10
   7.1. Normative References....10
   7.2. Informative References....10

8. Acknowledgments.....11
9. Contributing Authors.....11

1. Introduction

OAM (Operations, Administration and Maintenance) generally covers various production aspects of a network. In this document we use the term OAM as defined in [RFC6291].

Success of any mission critical network depends on the ability to proactively monitor networks for faults, performance, etc. as well as its ability to efficiently and quickly troubleshoot defects and failures. A well-defined OAM toolset is a vital requirement for wider adoption of TRILL as the next generation data forwarding technology in larger networks such as data centers.

In this document we define the Requirements for TRILL OAM. It is assumed that the readers are familiar with the OAM concepts and terminologies defined in other OAM standards such as [8021ag], [RFC5860]. This document does not attempt to redefine the terms and concepts specified elsewhere.

1.1. Scope

The scope of this document is OAM between RBridges of a TRILL campus over links selected by TRILL routing.

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]. Although this document is not a protocol specification, the use of this language clarifies the instructions to protocol designers producing solutions that satisfy the requirements set out in this document.

3. Terminology

Section: The term Section refers to a partial segment of a path between any two given RBridges. As an example, consider the case where RB1 is connected to RBx via RB2,RB3 and RB4. The segment between RB2 to RB4 is referred to as a Section of the path RB1 to RBx.

Flow: The term Flow indicates a set of packets that share the same path and per-hop behavior (such as priority). A flow is typically

identified by a portion of the inner payload that affects the hop-by hop forwarding decisions. This may contain Layer 2 through Layer 4 information.

All Selectable Least Cost Paths: The term "all selectable least cost paths" refers to a subset of all potentially available least cost paths to a specified destination RBridge that are available (and usable) for forwarding of frames. It is important to note, in practice, due to limitations in implementations, not all available least cost paths may be selectable for forwarding.

Connectivity: The term connectivity indicates reachability between an arbitrary RBridge RB1 and any other RBridge RB2. The specific path can be either explicit (i.e. associated with a specific flow) or unspecified. Unspecified means that messages used for connectivity verification take whatever that path the RBs happen to select.

Continuity Verification: Continuity Verification refers to proactive verification of Connectivity between two RBridges at periodic intervals and generation of explicit notification when Connectivity failures occur.

Fault: The term Fault refers to an inability to perform a required action, e.g., an unsuccessful attempt to deliver a packet.

Defect: The term Defect refers to an interruption in the normal operation, such that over a period of time no packets are delivered successfully.

Failure: The term Failure refers to the termination of the required function over a longer period of time. Persistence of a defect for a period of time is interpreted as a failure.

4. OAM Requirements

4.1. Data Plane

OAM frames, utilized for connectivity verification, continuity checks, performance measurements, etc., will by default take whatever the path TRILL chooses based on the current topology and per-hop equal cost path choices. In some cases, it may be required that the OAM frames utilize specific paths. Thus, it MUST be possible to arrange that OAM frames follow the path taken by a specific flow.

RBridges MUST have the ability to identify OAM frames destined for them or which require processing by the OAM plane from normal data frames.

TRILL OAM frames MUST NOT be forwarded out as native frames on end station service enabled ports.

OAM MUST have ability to include all Ethernet traffic types carried by TRILL, including both IP and non-IP traffic.

4.2. Connectivity Verification

4.2.1. Unicast

From an arbitrary RBridge RB1, OAM MUST have the ability to verify connectivity to any other RBridge RB2.

From an arbitrary RBridge RB1, OAM MUST have the ability to verify connectivity to any other RBridge RB2 for a specific flow via the path associated with the specified flow

An RBridge SHOULD have the ability to verify the above connectivity tests on sections. As an example, assume RB1 is connected to RB5 via RB2, RB3 and RB4. An operator SHOULD be able to verify the RB1 to RB5 connectivity on the section from RB3 to RB5. The difference is that the ingress and egress TRILL nicknames in this case are RB1 and RB5 as opposed to RB3 and RB5, even though the message itself may originate at RB3.

4.2.2. Multicast

OAM MUST have the ability to verify connectivity, from an arbitrary RBridge RB1, to either to specific set of RBridges or all member RBridges, for a specified multicast tree. This functionality is referred to as verification of the un-pruned multicast tree.

OAM MUST have the ability to verify connectivity, from an arbitrary RBridge RB1, to either to a specific set of RBridges or all member RBridges, for a specified multicast tree and for a specified flow. This functionality is referred to as verification of the pruned tree.

4.3. Continuity Check

OAM MUST provide functions that allow any arbitrary RBridge RB1 to perform a Continuity Check to any other RBridge.

OAM MUST provide functions that allow any arbitrary RBridge RB1 to perform a Continuity Check to any other RBridge using a path associated with a specified flow.

OAM SHOULD provide functions that allow any arbitrary RBridge to perform a Continuity Check to any other RBridge over all selectable least cost paths.

OAM SHOULD provide the ability to perform a Continuity Check on sections of any selectable path within the network.

OAM SHOULD provide the ability to perform a multicast Continuity Check for specified multi-destination tree(s) as well as specified multi-destination tree and flow combinations. The former is referred to as an un-pruned multi-destination tree Continuity Check and the latter is referred to as a pruned tree Continuity Check.

4.4. Path Tracing

OAM MUST provide the ability to trace a path between any two RBridges per specified unicast flow.

OAM SHOULD provide the ability to trace all selectable least cost paths between any two RBridges.

OAM SHOULD provide functionality to trace all branches of a specified multi-destination tree (un-pruned tree)

OAM SHOULD provide functionality to trace all branches of a specified multi-destination tree for a specified flow (pruned tree).

4.5. General Requirements

OAM MUST provide the ability to initiate and maintain multiple concurrent sessions for multiple OAM functions between any arbitrary RBridge RB1 to any other RBridge. In general, multiple OAM operations will run concurrently. For example, proactive continuity checks may take place between RB1 and RB2 at the same time an operator decides to test connectivity between the same two RBs. Multiple OAM functions and instances of those functions MUST be able to run concurrently without interfering with each other.

OAM MUST provide a single OAM framework for all TRILL OAM functions within the scope of this document.

OAM, as practical and as possible, SHOULD provide a single framework between TRILL and other similar standards.

OAM MUST maintain related error and operational counters. Such counters MUST be accessible via network management applications (e.g. SNMP).

OAM functions related to continuity and connectivity checks MUST be able to be invoked either proactively or on-demand.

OAM SHOULD NOT require extensions to the TRILL header. OAM MAY be required to provide the ability to specify a desired response mode for a specific OAM message. The desired response mode can be either in-band, out-of band or none.

The OAM Framework MUST be extensible to future needs of TRILL and the needs of other standard organizations.

OAM MAY provide methods to verify control plane and forwarding plane alignments.

OAM SHOULD leverage existing OAM technologies, where practical.

4.6. Performance Monitoring

4.6.1. Packet Loss

In this document, term loss of a packet is used as defined in [RFC2680] (see Section 2.4 of RFC2680).

NOTE: The term simulated flow below indicates a flow that is generated by an RBRidge RB1 for OAM purposes. The fields of the simulated flow may or may not be identical to the actual data. However, simulated flow is required to follow the intended path.

OAM SHOULD provide the ability to measure packet loss statistics for a simulated flow from any arbitrary RBridge RB1 to any other RBridge.

OAM SHOULD provide the ability to measure packet loss statistics over a segment, for a simulated flow between any arbitrary RBridge RB1 to any other RBridge.

OAM SHOULD provide the ability to measure simulated packet loss statistics between any two RBridges over all least cost paths.

An RBridge SHOULD be able to perform the above packet loss measurement functions either proactively or on-demand.

4.6.2. Packet Delay

There are two types of packet delays -- one-way delay and two-way delay (Round Trip Delay).

One-way delay is defined in [RFC2679] as the time elapsed from the start of transmission of the first bit of a packet by an RBridge until the reception of the last bit of the packet by the destination RBridge.

Two-way delay is also referred to as Round Trip Delay is defined similar to [RFC2681]; i.e. the time elapsed from the start of transmission of the first bit of a packet by an RBridge until the reception of the last bit of the packet by the same RBridge.

OAM SHOULD provide functions to measure two-way delay between two RBridges for a specified flow.

OAM SHOULD provide functions to measure two-way delay between two RBridges for a specified flow over a specific section.

OAM MAY provide functions to measure one-way delay between two RBridges for a specified flow.

OAM MAY provide functions to measure one-way delay between two RBridges for a specified flow over a specific section.

4.7. ECMP Utilization

OAM MAY provide functionality to monitor the effectiveness of per-hop ECMP hashing. For example, individual RBridges could maintain counters that show how packets are being distributed across equal cost next hops for a specified destination RBridge or RBridges as a result of ECMP hashing.

4.8. Security and Operational considerations

Methods MUST be provided to protect against exploitation of OAM framework for security and denial of service attacks.

Methods SHOULD be provided to prevent OAM messages causing congestion in the networks. Periodically generated messages with high frequencies may lead to congestion, hence methods such as shaping or rate limiting SHOULD be utilized.

4.9. Fault Indications

The term Fault refers to an inability to perform a required action, e.g., an unsuccessful attempt to deliver a packet [OAMOVER]. The unsuccessful attempt may be due to Hop Count expiry, invalid nickname, etc.

OAM MUST provide a Fault Indication framework to notify faults to the ingress RBRidge of the packet or other interested parties (such as syslog servers).

OAM MUST provide functions to selectively enable or disable different types of Fault Indications.

4.10. Defect Indications

[OAMOVER] defines "The term Defect refers to an interruption in the normal operation, such as a consecutive period of time where no packets are delivered successfully."

OAM SHOULD provide a framework for Defect Detection and Indication.

OAM implementations that provide Defect Indication MUST provide methods to selectively enable or disable Defect Detection per defect type.

OAM implementations that provide Defect Indication MUST provide methods to configure Defect Detection thresholds per different types of defects.

OAM implementations that provide Defect Indication facilities MUST provide methods to log defect indications to a locally defined archive such as log buffer or SNMP traps.

OAM implementations that provide Defect Indication facilities SHOULD provide a Remote Defect Indication framework that facilitates notifying the originator/owner of the flow experiencing the defect, which is the ingress RBridge.

Remote Defect Indication MAY be either in-band or out-of-band.

4.11. Live Traffic monitoring

OAM implementations MAY provide methods to utilize live traffic for troubleshooting and performance monitoring.

Implementations MAY leverage Data Driven CFM [8021Q] or IPFIX [RFC5101] for the purpose of performance monitoring.

5. Security Considerations

Security Requirements are specified in section 4.8. For general TRILL security considerations please refer to [RFC6325]

6. IANA Considerations

None

7. References

7.1. Normative References

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

7.2. Informative References

[RFC6325] Perlman, R., et.al., "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC5101] Claise, B., "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information", RFC5101, January 2008.

[RFC2680] Almes, G., et.al. "A One-way Packet Loss Metric for IPPM", RFC 2680, September 1999.

[RFC2679] Almes, G., et.al. "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC2681] Almes, G., et.al. "A Round-trip Delay Metric for IPPM", RFC 2681, September 1999.

[RFC6291] Anderson, L., et.al. "Guidelines for the Use of the "OAM" Acronym in the IETF", RFC 6291, June 2011.

[8021ag] IEEE, "Virtual Bridged Local Area Networks Amendment 5: Connectivity Fault Management", 802.1ag, 2007.

[8021Q] IEEE, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, August, 2011.

[RFC4377] Nadeau, T., et.al. "Operations and Management (OAM) Requirements for Multi-protocol Label Switched (MPLS)Networks", RFC 4377, February 2006.

[OAMOVER] Mizrahi, T, et.al., "An Overview of Operations, Administration, and Maintenance (OAM) Mechanisms", draft-ietf-opsawg-oam-overview-06, Work in Progress, March 2012.

[RFC5860] Vigoureux, M., et.al., "Requirements for Operations, Administration and Maintenance (OAM) in MPLS Transport Networks", RFC5860, May 2010.

8. Acknowledgments

Special acknowledgments to IEEE 802.1 chair, Tony Jeffree for allowing us to solicit comments from IEEE 802.1 group. Also recognized are the comments received from IEEE group, Ayal Lior and others.

This document was prepared using 2-Word-v2.0.template.dot.

9. Contributing Authors

Tissa Senevirathne
CISCO Systems
375 East Tasman Drive
San Jose, CA 95134
USA.

Phone: +1-408-853-2291
Email: tsenevir@cisco.com

David Bond
IBM
2051 Mission College Blvd
Santa Clara, CA 95054
USA

Phone: +1-603-339-7575
Email: mokon@mokon.net

Sam Aldrin
Huawei Technologies
2330 Central Express Way
Santa Clara, CA 95951
USA

Email: aldrin.ietf@gmail.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

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

Rohit Watve
CISCO Systems
375 East Tasman Drive
San Jose, CA 95134
USA.

Phone: +1-408-424-2091
Email: rwatve@cisco.com

Thomas Narten
IBM Corporation
3039 Cornwallis Avenue,
PO Box 12195
Research Triangle Park, NC 27709
USA

Email:narten@us.ibm.com

Donald Eastlake
Huawei Technologies
155 Beaver Street,
Milford, MAC 01757
USA.

Email: d3e3e3@gmail.com

Anoop Ghanwani
DELL
350 Holger Way
San Jose, CA 95134
USA.

Phone: +1-408-571-3500
Email: Anoop@alumni.duke.edu

Jon Hudson
Brocade
120 Holger Way
San Jose, CA 95134
USA.

Email: jon.hudson@gmail.com

Naveen Nimmu
Broadcom
9th Floor, Building no 9, Raheja Mind space
Hi-Tec City, Madhapur,
Hyderabad - 500 081, INDIA

Phone: +1-408-218-8893
Email: naveen@broadcom.com

Radia Perlman
Intel Labs
2700 156th Ave NE, Suite 300,
Bellevue, WA 98007
USA.

Phone: +1-425-881-4824
Email: radia.perlman@intel.com

Tal Mizrahi
Marvell
6 Hamada St.
Yokneam, 20692 Israel

Email: talmi@marvell.com

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: March 25, 2013

Kesava Vijaya Krupakaran
Janardhanan Pathangi Narasimhan
Dell

September 21, 2012

Fair Share AF Load Share
draft-kvk-trill-fair-share-af-load-share-02

Abstract

In an access LAN of a TRILL campus, the DRB can choose to load share the AF responsibility among other RBridges in the LAN. This document throws light on one such approach where the AF appointment is fair share scheduled among the RBridges.

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

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1 Introduction 3
1.1 Terminology 3
2 Shares 3
3 AF Affinity VLAN Set 4
4 AF Affinity VLAN Set Overlap 5
5 AF Distribution Among Heterogeneous RBridges 6
6 AF Computation at DRB 6
7 AF and VLAN Mapping 7
8 AF and Multiple ports on a link 7
9 Multi-Topology-Aware Port Capability Sub-TLVs 7
9.1 Fair Share Sub-TLV 7
9.2 AF Affinity VLAN Set Sub-TLV 7
9.3 Partial VLANs Appointing Sub-TLV 8
10 Security Considerations 9
11 IANA Considerations 9
12 References 9
12.1 Normative References 9
12.2 Informative References 9
Authors' Addresses 10

1 Introduction

In a shared access LAN, the appointed forwarder for a VLAN is responsible for encapsulating and decapsulating native traffic on that VLAN. Other non-AF RBridges in the LAN discard the native traffic for that VLAN.

The DRB can choose to be the AF for all VLANs or load share the AF responsibility among other RBRidges in the LAN. This ensures better utilization of resources like hardware tables and buffers. The VLAN partitioning scheme suggested in [RFC6439] section 2.2.1 is static and requires careful configuration. Another simple protocol would be to allocate VLANs in a round-robin fashion among all RBridges in the LAN. However this doesn't leave scope for schemes like retaining 50% of VLANs with the DRB and distribute only the rest among others.

Fair share scheduling of AF allows for the flexibility of assigning certain RBridges (say with higher switching capability) AF for higher proportion of VLANs than others.

1.1 Terminology

This document uses the acronyms defined in [RFC6439].

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].

2 Shares

Each RBridge is configured with certain quantity of shares. A share is the proportion of VLANs which would be allocated to the RBridge in comparison with other RBridges. The face value of the shares is a relative quantity and makes sense only when taken in conjunction with total shares allocated in the LAN.

These shares are advertised by each RBridge in its hello. The DRB load shares the AF among RBridges based on the relative value of shares.

For instance, let A, B and C be three RBridges with S(A) = 2, S(B) = 1 and S(C) = 1. Then A is assigned the AF for 1/2 of the VLANs while B and C the AF for 1/4th of the VLANs each.

Even when the number of VLANs for which the RBridge is to be AF calculates to a non integer value, it should be made sure that there

is only one AF for a VLAN in a multi-access LAN.

3 AF Affinity VLAN Set

Fair share scheduling distributes VLANs among RBridges according to proportion of shares allocated. This allows allocation of higher proportion of VLANs to certain RBridges (with higher switching capability). However, this does not guarantee that these RBridges would handle larger share of the native traffic.

Following the previous example, even though A is appointed AF for 50% of the VLANs while B only 25% of the VLANs, the traffic load of VLANs for which B is AF could be considerably higher that those in A.

In order to overcome this conundrum, each RBriged in access LAN is configured with an AF Affinity VLAN Set apart form the share proportion. This RBridge has AF affinity to the set of configured VLANs. Thus when the DRB appoints an RBridge AF for a set of VLANs, the members of the set are chosen from the AF Affinity VLAN Set advertised.

Expanding on the previous example, if X denotes an RBridge, let S(X) be the shares assigned to X, V(X) be the AF Affinity VLAN Set and AF(X) denote the set of VLANs for which X is assigned AF. Let the access LAN encompass ten shared VLANs [11, 20]. In this case the AF assignment with just the shares configured could be as in Table 1. If RBridge A has higher switching capability and VLANs [16, 20] are heavily loaded, this AF appointment defeats the purpose.

Table 1: AF appointment using fair share scheduling

X	S(X)	AF(X)
A	2	{11, 12, 13, 14, 15}
B	1	{16, 17, 18}
C	1	{19, 20}

By configuring AF Affinity VLAN set in each RBridge, this difficulty can be overcome. Such a configuration is shown in Table 2. How the AF Affinity VLAN set is arrived at is beyond the scope of this document. Long term traffic planning tools could be helpful in extrapolating a decent configuration.

Table 2: Fair share scheduling with AF Affinity VLAN set			
X	S(X)	V(X)	AF(X)
A	2	{16, 17, 18, 19, 20}	{16, 17, 18, 19, 20}
B	1	{11, 12, 13, 14}	{11, 12, 13}
C	1	{12, 13, 14}	{14, 15}

4 AF Affinity VLAN Set Overlap

If the AF Affinity VLAN sets advertised by the RBridges overlap, the RBridge with higher share has priority over the affinity of common VLANs. In case the RBridges advertise same share with conflicting AF Affinity VLAN sets, then the one with higher system ID gets more AF affinity over the common VLANs.

Table 3: Fair share scheduling with AF Affinity VLAN set overlap				
X	ID(X)	S(X)	V(X)	AF(X)
A	0000.0000.000a	2	{15, 16, 17, 18, 19, 20}	{15, 16, 17, 18, 19}
B	0000.0000.000b	1	{11, 12, 13, 14, 15}	{14, 20}
C	0000.0000.000c	1	{11, 12, 13, 14}	{11, 12, 13}

Consider the previous examples with the LAN comprising of three RBridges A, B and C coloured for VLANs [11, 20]. As shown in table 3, the AF Affinity VLAN sets overlap in RBridges {A, C} as well as {B, C}. A, having the highest share has the most affinity over the VLANs configured there. In this example, A has higher AF affinity to VLAN 15 than C. Similarly, C has greater the AF affinity of VLANs [11, 14] than B on virtue of its higher system ID.

It is possible to calculate a better AF distribution by examining common VLANs in AF Affinity VLAN sets when they overlap. Such algorithms have been avoided to keep the computation at DRB simple.

5 AF Distribution Among Heterogeneous RBridges

An access LAN could constitute motley set of RBridges with some that support fair share AF scheduling and some that doesn't. If the DRB doesn't support fair share AF scheduling, it ignores the sub-TLVs advertised by other RBridges and continue to distribute AF as it did previously.

If the DRB does support fair share AF scheduling and it receives hello form an RBridge without Fair Share Sub-TLV, it is assigned a default share equal to average of all shares advertised in the LAN during AF computation. If the AF Affinity VLAN Set Sub-TLV was not advertised, it is taken to be a NULL set. In case an RBridge advertises AF Affinity Sub-TLV without saying the shares, such TLV is ignored and the behaviour follows as though it had not advertised AF Affinity VLAN set.

In particular, if DRB is the only RBridge supporting the feature, all the RBridges get equal shares (equal to the one configured at DRB, consistent with the average rule discussed).

For instance, in an access LAN with RBridges A, B and C where S(A) = 7, S(B) = 3, and C doesn't support fair share AF scheduling, the DRB assigns it a default of 5 shares.

As a special case, if DRB supports fair share AF load share and none of the RBridges advertise any share and no share is configured in DRB, then DRB assigns a share value of 1 to all RBridges and load shares VLANs equally among all the RBridges.

6 AF Computation at DRB

DRB runs through all RBridges, in descending order of shares configured and assigns the AF based on Affinity VLAN set. If the shares advertised are equal, then RBridges are ordered based on system ID. If RBridges don't advertise shares, they are assigned default shares and are placed below RBridges who advertise shares in the ordered list of RBridges. If there are multiple such non congruous RBridges, they are again ordered based on system ID.

The DRB also monitors hellos for any change from previously advertised shares or AF Affinity VLAN set. If it detects a change, the AF assignment is recomputed for all RBridges. Any addition or deletion of adjacency also triggers fresh AF assignment. This simplifies the computation at DRB.

7 AF and VLAN Mapping

If the DRB detects VLAN mapping, it appoints one RBridge (possibly itself) as the AF for all VLANs as suggested in [RFC6439] section 2.4 to prevent loops.

8 AF and Multiple ports on a link

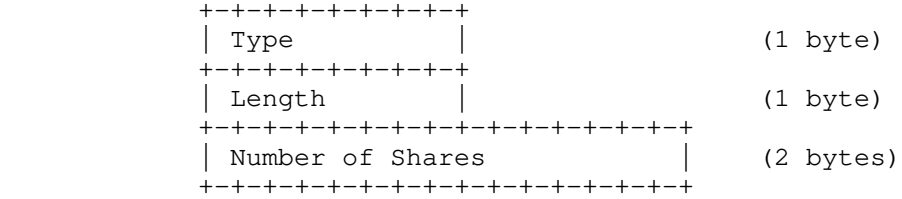
The shares configured represents the whole RBridge's proportion of AF sought. Further load sharing of AF among multiple ports on same link in an RBridge is a local decision.

9 Multi-Topology-Aware Port Capability Sub-TLVs

Two new Multi-Topology-Aware Port Capability Sub-TLVs are required for the purpose of fair share AF appointment - Fair Share Sub-TLV and AF Affinity VLAN Set Sub-TLV.

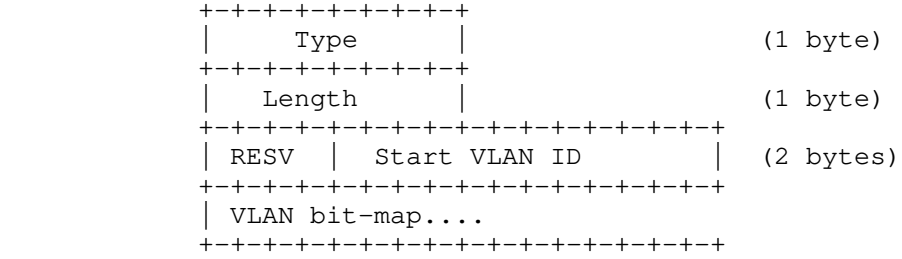
9.1 Fair Share Sub-TLV

Fair Share Sub-TLV is used to advertise the number of shares configured in the RBridge. Number of shares is a two octet value. When an RBridge advertises zero shares, it is not assigned any AF.



9.2 AF Affinity VLAN Set Sub-TLV

AF Affinity VLAN Set Sub-TLV is used to advertise the AF Affinity VLAN set configured in an RBridge. It is a facsimile of the Enabled-VLANs sub-TLV.



9.3 Partial VLANs Appointing Sub-TLV

As discussed in [RFC6439] section 2.2.3, the size of hello imposes a limit on the distribution of AF info in AF Sub-TLV by the DRB. The nature of the algorithm means that the AF appointment information could be disjoint. If the number of VLANs on a shared link is too high, all AF appointments cannot be accommodated in a single hello using the start end mechanism of AF Sub-TLV. In such case, the DRB should appoint one RBridge (possibly itself) as AF for all VLANs.

Alternatively, the AF information can be sent in a bitmap rather than start-end mechanism as suggested in AF Sub-TLV. For this purpose the Partial VLANs Appointing Sub-TLV suggested in Adaptive VLAN Assignment draft [VlanAsn] can be used.

10 Security Considerations

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

11 IANA Considerations

This document suggests two additional Sub-TLV to Multi-Topology-Aware Port Capability TLV apart from the reuse of Partial VLANs Appointing Sub-TLV from Adaptive VLAN Assignment draft.

- Fair Share Sub-TLV
- AF Affinity VLAN Set Sub-TLV

12 References

12.1 Normative References

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

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

[RFC6326] D. Eastlake, A. Banerjee, et al, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 6326, July 2011.

[RFC6439] D. Eastlake, R. Perlman, et al, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.

[RBisisb] D. Eastlake, A. Banerjee, et al, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-09.txt, work in progress.

12.2 Informative References

[VlanAsn] M.Zhang and D.Zhang, "Adaptive VLAN Assignment for Data Center RBridges", draft-zhang-trill-vlan-assign-04.txt, work in progress.

Authors' Addresses

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

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

Janardhanan Pathangi
Dell
Olympia Technology Park,
Guindy Chennai 600 032

Phone: +91 44 4220 8459
Email: Pathangi_Janardhanan@Dell.com

TRILL Working Group
Internet Draft
Intended status: Standard Track

Tissa Senevirathne
Samer Salam
Deepak Kumar
CISCO

Donald Eastlake
Sam Aldrin
YiZhou Li
Huawei

September 7, 2012

Expires: March 2013

TRILL Fault Management
draft-tissa-trill-oam-fm-00.txt

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), 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/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on March 7, 2009.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

In this document we present definitions of TRILL OAM messages. Messages defined in this document follow a similar structure to IEEE 802.1ag messages.

Table of Contents

1. Introduction....3
2. Conventions used in this document....4
3. General Format of TRILL OAM frames....4
3.1. Identification of TRILL OAM frames....6
3.2. Use of TRILL OAM Flag....6
3.2.1. Handling of TRILL frames with F flag....7
3.3. Backwards Compatibility Method....8
3.4. OAM Capability Announcement....8
4. TRILL OAM Message Channel....9
4.1. TRILL OAM Message header....9
4.2. TRILL OAM Opcodes....10
4.3. Format of TRILL OAM TLV....11
4.4. TRILL OAM TLVs....12
4.4.1. Common TLVs between 802.1ag and TRILL....12
4.4.2. TRILL OAM Specific TLVs....12
4.4.2.1. TRILL OAM Application Identifier TLV....12
4.4.3. Out Of Band IP Address TLV....14
4.4.3.1. Diagnostics VLAN TLV....14
4.4.3.2. Original Data Payload TLV....15
4.4.3.3. RBridge scope TLV....15
4.4.3.4. Upstream RBridge nickname TLV....16
4.4.3.5. Next Hop RBridge List TLV....17
4.4.3.6. Multicast Receiver Availability TLV....17
5. Loopback Message....18
5.1.1. Loopback OAM Message format....18
5.1.2. Theory of Operation....19
5.1.2.1. Originator RBridge....19
5.1.2.2. Intermediate RBridge....19
5.1.2.3. Destination RBridge....19
6. Path Trace Message....20

6.1.1. Theory of Operation.....21
6.1.1.1. Originator RBridge.....21
6.1.1.2. Intermediate RBridge.....21
6.1.1.3. Destination RBridge.....22
7. Multicast Tree Verification (MTV) Message.....22
7.1. Multicast Tree Verification (MTV) OAM Message Format.....23
7.2. Theory of Operation.....23
7.2.1. Originator RBridge.....23
7.2.2. Receiving RBridge.....24
7.2.3. In scope RBridges.....25
8. Notification Messages.....25
9. Return Codes.....26
9.1. Return Codes.....26
9.2. Return sub-codes.....26
10. Security Considerations.....27
11. Allocation Considerations.....27
12. References.....27
12.1. Normative References.....27
12.2. Informative References.....27
13. Acknowledgments.....28

1. Introduction

The general structure of TRILL OAM messages is presented in [TRLOAMFRM]. According to [TRLOAMFRM], TRILL OAM messages consist of four main parts: link header, TRILL header, flow entropy, and OAM message channel.

The OAM message channel allows defining various control information and carrying OAM related data between TRILL switches, also known as RBridges or Routing Bridges.

The OAM message channel, if defined properly, can be shared between different technologies. A common OAM channel allows a uniform user experience for the customers, savings on operator training, re-use of software code base, and faster time to market.

This document uses the message format defined in IEEE 802.1ag Connectivity Fault Management (CFM) [802.1ag] [802.1Q] as the basis for the TRILL OAM channel message.

The ITU-T Y.1731 standard utilizes the same messaging format as [802.1ag]. However, IEEE defines a separate op-code space for the messages specific to Y.1731. This document specifies asimilar approach for TRILL and request a separate code space to be assigned for TRILL OAM messages.

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].

Acronyms used in the document include the following:

MP - Maintenance Point [TRLOAMFRM]

OAM - Operations, Administration, and Maintenance [RFC6291]

TRILL - Transparent Interconnection of Lots of Links [RFC6325]

3. General Format of TRILL OAM frames

The TRILL forwarding paradigm allows an implementation to select a path from a set of equal cost paths to forward a packet. Selection of the path of choice is implementation dependent. However, it is a common practice to utilize Layer 2 through Layer 4 information in the frame payload for path selection.

For accurate monitoring and/or diagnostics, OAM Messages are required to follow the exact path as the data packets. [TRILLOAMFM] proposes a high-level format of the OAM messages. The details of the TRILL OAM frame format are defined in this document.

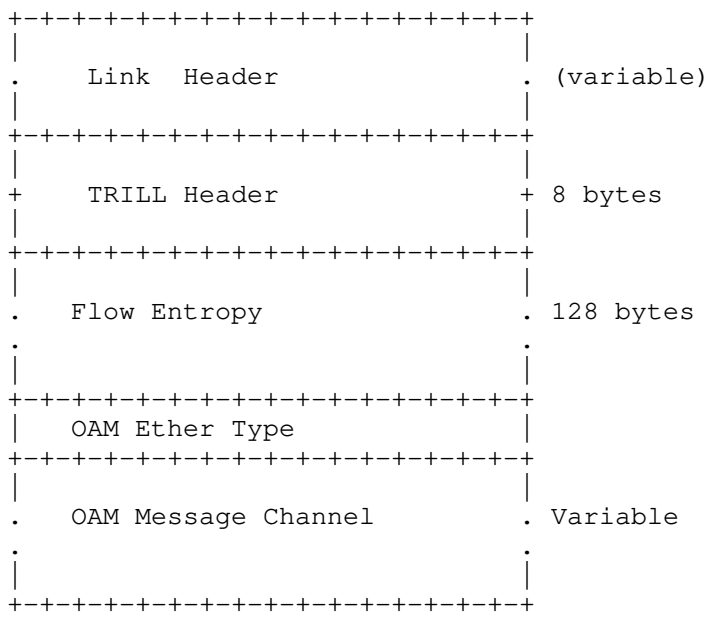


Figure 1 Frame format of OAM Messages

Link Header: Media dependent header. For Ethernet this included Destination MAC, Source MAC, VLAN (optional) and EtherType fields.

TRILL Header: Minimum of 8 bytes when the Extended Header is not included [RFC6325]

Flow Entropy: This is a 128-byte Fixed size opaque field. The least significant bits of the field MUST be padded with zeros up to 128 bytes, when the flow entropy is less than 128 bytes. Flow entropy enables emulation of the forwarding behavior of the desired data packets.

OAM Ether Type: OAM Ether Type is 16-bit EtherType that identifies the OAM Message channel that follows. This document specifies to use EtherType allocated for 802.1ag for the purpose. Identifying the OAM Message Channel with a dedicated EtherType allows the easy identification of the beginning of the OAM message channel across multiple Ethernet standards.

OAM Message Channel: This is a variable size section that carries OAM related information. Reusing existing OAM message definitions such as [RFC4379] and [8021ag] will be explored.

3.1. Identification of TRILL OAM frames

TRILL, as defined in [RFC6325], does not have a specific flag or a method to identify OAM related frames. This document specifies to update RFC6325 to include specific methods to identify TRILL OAM frames. Section 3.2. 3.2. below explains the details of the method. However, it is important, for backwards compatibility reasons, to define methods to identify TRILL OAM frames without using the extensions. Section 3.3. presents a set of possible methods for identifying OAM frames without using the proposed extensions in section 3.2. Methods defined in section 3.3. impose limitations on the construction of the flow entropy of the OAM frames and MUST be used for backwards compatibility scenarios only.

3.2. Use of TRILL OAM Flag

The TRILL Header, as defined in [RFC6325], has two reserved bits that are currently unused. RBridges are currently required to ignore these fields. This document specifies to use the reserved bit next to Version field in the TRILL header as the OAM flag. OAM flag will be denoted by 'F'.

Implementations that follow the extension of using the F flag to identify TRILL OAM frames MUST exclusively use that flag as the means of identifying OAM frames, as specified in section 3.2.1. The F flag MUST NOT be utilized for other forwarding decisions such as selection of ECMP paths etc.

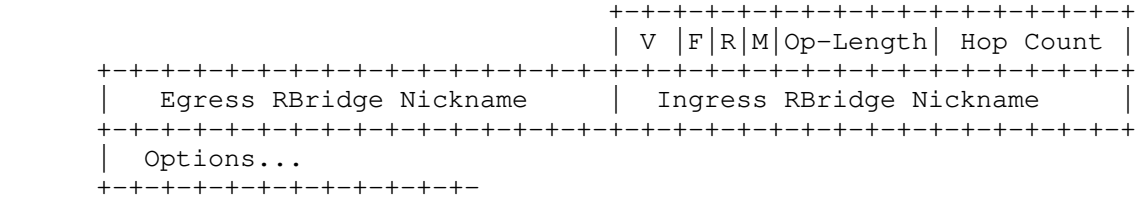


Figure 2 TRILL Header

F (1 bit) - Indicates this is an OAM frame and is subject to specific handling as specified in this document.

All other fields carry the same meaning as defined in RFC6325.

3.2.1. Handling of TRILL frames with F flag

When a unicast TRILL encapsulated frame is received with the F flag set, the following processing occurs:

If the egress RBridge nickname is local, the frame MUST be forwarded to the CPU for further processing and MUST NOT be forwarded out of the RBridge.

If the egress RBridge nickname is not local, the frame MUST be forwarded as specified in [RFC6325].

When a multicast TRILL encapsulated frame is received with the F flag set, the following processing occurs:

A copy of the frame MUST be sent to the CPU for further processing.

Additionally, the frame MUST also be forwarded as specified in [RFC6325].

A TRILL encapsulated frame with the F flag set MUST NOT be de-capsulated and forwarded as a native frame.

```
Receiver Processing:

If (M==1 && F==1) then
   Copy to CPU and Forward normally as defined in RFC 6325
Else if (M==0 && F==1 && egress nickname is the processing RBridge)
then
   Forward to CPU BUT DO NOT forward along the data plane

Else
   Forward as defined in RFC 6325
End;

Transmit Processing:

If (F==1) then
  Forward as defined in RFC6325 BUT Do not de-capsulate and forward
as a native frame
Else
  Forward as defined in RFC 6325
```

Figure 3 Pseudo code for F flag processing

3.3. Backwards Compatibility Method

For unicast frames, TRILL MP is addressed by its TRILL egress nickname and either OAM Inner.MacSA or OAM Ethtype .

For unicast frames, a TRILL MP (Maintenance Point) is addressed by combination of TRILL nickname of the target TRILL RBridge (where the MP resides as the egress nickname of the TRILL Header in the TRILL OAM frame) and either the OAM Inner.MacSA or the OAM Ethertype

For multicast frames, TRILL MP is addressed by either Reserved EtherType or Reserved source MAC .

The following table summarizes the identification of different OAM frames from data frames.

Flow Entropy	Inner MacSA	OAM Eth Type	Egress nickname
unicast L2	N/A	Match	Match
Multicast L2	N/A	Match	N/A
Unicast IP	Match	N/A	Match
Multicast IP	Match	N/A	N/A
Notification	N/A	Match	Match

Figure 4 Identification of TRILL OAM Frames

3.4. OAM Capability Announcement

Any given TRILL RBridge can be one of: OAM incapable OR OAM capable with new extensions OR OAM capable with backwards compatible method. The OAM request originator, prior to origination of the request is required to identify the OAM capability of the target and generate the appropriate OAM message.

We propose to utilize capability flags defined in TRILL version sub-TLV (TRILL-VER) [rfc6326bis]. The following Flags are defined:

O - OAM Capable

B - Backwards Compatible.

A capability announcement, with O Flag set to 1 and B flag set to 1, indicates that the implementation is OAM capable but utilize backwards compatible method defined in section Error! Reference source not found.Error! Reference source not found.

A capability announcement, with O Flag set to 1 and B flag set to 0, indicates that the implementation is OAM capable and utilizes the method specified in section 3.2.

When O Flag is set to 0, the announcing implementation is considered not capable of OAM and B flag is ignored.

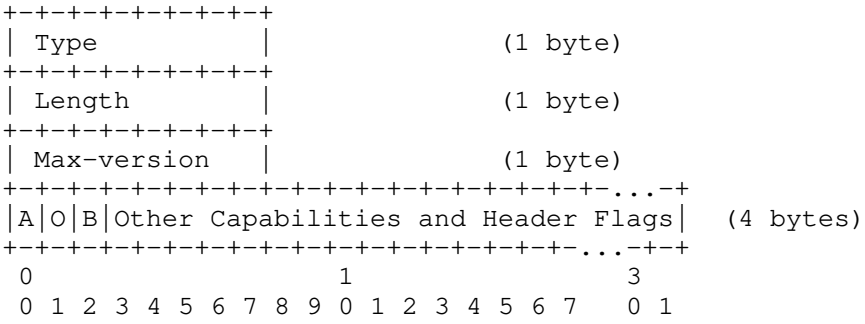


Figure 5 TRILL-VER sub-TLV [rfc6326bis] with O and B flags

4. TRILL OAM Message Channel

The TRILL OAM Message Channel can be divided in to two parts: TRILL OAM Message header and TRILL OAM Message TLVs. Every OAM Message MUST contain a single TRILL OAM message header and a set of one or more specified OAM Message TLVs.

4.1. TRILL OAM Message header

As discussed earlier, we propose to use the Message format defined in IEEE 802.1ag. We believe a common messaging framework between [802.1ag], TRILL and other similar standards such as Y.1731 can be accomplished by re-using the OAM message header defined in [802.1ag] and [802.1Q].

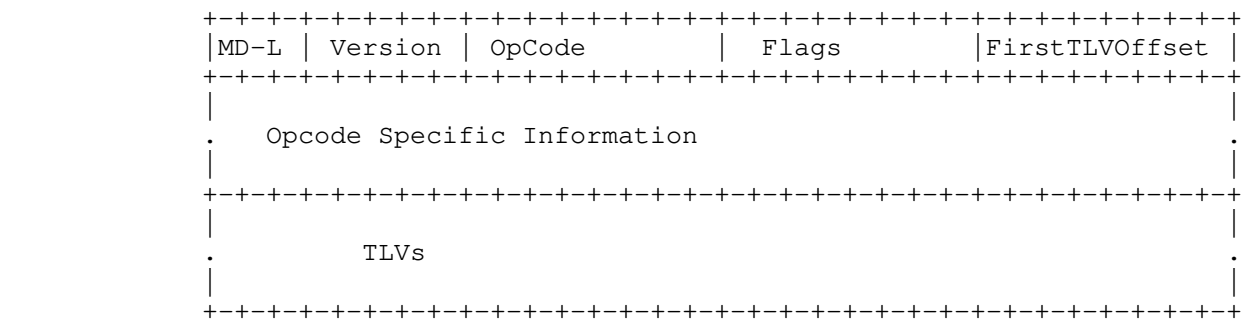


Figure 6 OAM Message Format

- MD-L: Maintenance Domain Level (3 bits). Identifies the maintenance domain level. For TRILL this MAY be always set to zero. However, in multilevel TRILL [TRILLML], backbone MAY be of a different MD-LEVEL. (Please refer to [802.1ag] for the definition of MD-Level)

- Version: Indicates the version (5 bits). [802.1ag] sets the version to zero.

- Flags: Include operational flags (1 byte). The definition of flags is Opcode specific and is covered in the applicable sections.

- FirstTLVOffset: Defines the location of the first TLV, in bytes, starting from the end of the FirstTLVOffset field (1 byte). (Refer to [802.1ag] for the definition of the FirstTLVOffset.)

MD-L, Version, Opcode, Flags, FirstTLVOffset, fields collectively are referred to as the OAM Message Header.

The Opcode specific information section of the OAM Message may contain Session Identification number, time-stamp, etc. Details about the Opcode specific information section and the associated TLVs will be presented later in this document.

4.2. TRILL OAM Opcodes

Following Opcodes are defined for TRILL. Each of the opcodes defines a separate TRILL OAM message. Details of the messages are presented in the related sections.

TRILL OAM Message Opcodes:

64 : Loopback Message Reply
65 : Loopback Message
66 : Path Trace Reply
67 : Path Trace Message
68 : Notification Message
69 : Multicast Tree Verification Reply
70 : Multicast Tree Verification Message
71 : Performance Measurement one-way Reply
72 : Performance Measurement one-way Message
73 : Performance Measurement two-way Reply
74 : Performance Measurement two-way Message
75 - 95 : Reserved

4.3. Format of TRILL OAM TLV

We propose to use the same format as defined in section 21.5.1 of [802.1ag]. The following figure depicts the general format of TRILL OAM TLV:

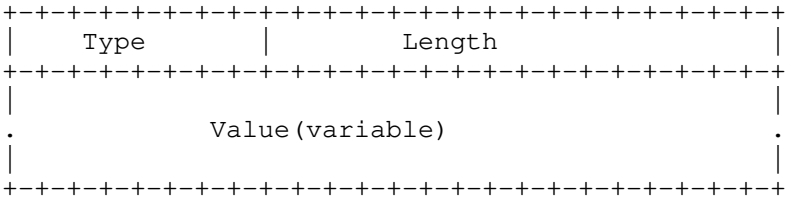


Figure 7 TRILL OAM TLV

Type (1 octet) : Specifies the Type of the TLV (see sections 4.4. 4.4.1. 4.4.2. for TLV types).

Length (2 octets) : Specifies the length of the values field in octets. Length of the value field can be either zero or more octets.

Value (variable): Length and the content of the value field depend on the type of the TLV. Please refer to applicable TLV definitions for the details.

Semantics and usage of Type values allocated for the TRILL OAM purpose are defined by this document and other future related documents.

4.4. TRILL OAM TLVs

In this section we define TRILL related TLVs. We propose to re-use [802.1ag] defined TLVs where applicable. Types 32-63 are reserved for ITU-T Y.1731. We propose to reserve Types 64-95 for the purpose of TRILL OAM TLVs.

4.4.1. Common TLVs between 802.1ag and TRILL

Following TLVs are defined in [802.1ag], we propose to re-use them where applicable. Format and semantics of the TLVs are as defined in [802.1g]. NOTE: Presented within brackets is the corresponding Type.

1. End TLV (0)
2. Sender ID TLV (1)
3. Port Status TLV (2)
4. Data TLV (3)
5. Interface Status TLV (4)
6. Reply Ingress TLV (5)
7. Reply Egress TLV (6)
8. LTM Egress Identifier TLV (7)
9. LTR Egress Identifier TLV (8)
10. Reserved (9-30)
11. Organization specific TLV (31)

4.4.2. TRILL OAM Specific TLVs

As indicated above, Types 64-95 will be requested to be reserved for TRILL OAM purposes. Listed below, a summary of TRILL OAM TLV and the corresponding codes. Format and semantics of TRILL OAM TLVs are defined in subsequent sections.

1. TRILL OAM Application Identifier (64)
2. Out of Band IP Address (65)
3. Diagnostic VLAN (66)
4. RBridge scope (67)
5. Original Payload (68)
6. Upstream RBridge nickname(69)
7. TRILL Next Hop RBridge List (ECMP) (70)
8. Multicast Receiver Availability (71)
9. Reserved (71-95)

4.4.2.1. TRILL OAM Application Identifier TLV

TRILL OAM Application Identifier TLV carries TRILL OAM application specific information. The TRILL OAM Application Identifier TLV MUST always be present and MUST be the first TLV in TRILL OAM messages. Messages that do not include the TRILL OAM Application Identifier TLV as the first TLV MUST be discarded.

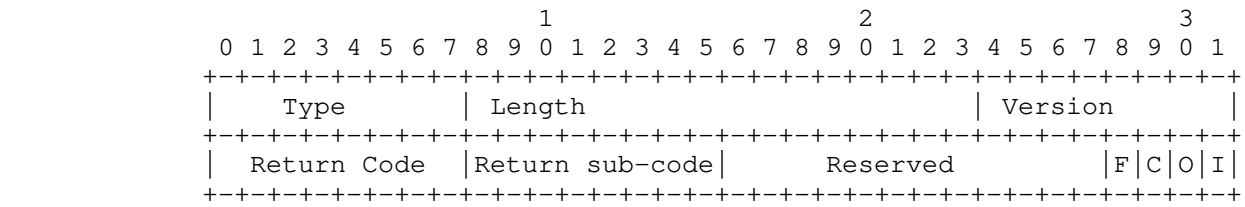


Figure 8 TRILL OAM Message TLV

Type (1 octet) = 64 indicate that this is the TRILL OAM Version

Length (2 octets) = 6

Version (1 Octet), currently set to zero. Indicates the TRILL OAM version. TRILL OAM version can be different than the [802.1ag] version.

Return Code (1 Octet): Set to zero on requests. Set to an appropriate value in response or notification messages. Please see section x below for definition of return codes.

Return sub-code (1 Octet): Return sub-code is set to zero on transmission of request message. Return sub-code identifies categories within a specific Return code. Return sub-code MUST be interpreted within a Return code and specified in section x below.

Reserved: set to zero on transmission and ignored on reception.

F (1 bit) : Final flag, when set, indicates this is the last response.

C (1 bit): Cross connect error (VLAN mapping error), if set indicates VLAN cross connect error detected. This field is ignored in request messages and MUST only be interpreted in response messages.

O (1 bit) : If set, indicates, OAM out-of-band response requested.

I (1 bit) : If set, indicates, OAM in-band response requested.

NOTE: When both O and I bits are set to zero, indicates that no response is required (silent mode). User MAY specify both O and I or one of them or none.

4.4.3. Out Of Band IP Address TLV

Out of Band IP Address TLV specifies the IP address to which an out of band OAM reply message MUST be sent. When O bit in the Version TLV is not set, Out of Band IP Address TLV is ignored. Length of the TLV implies the IP Address version.

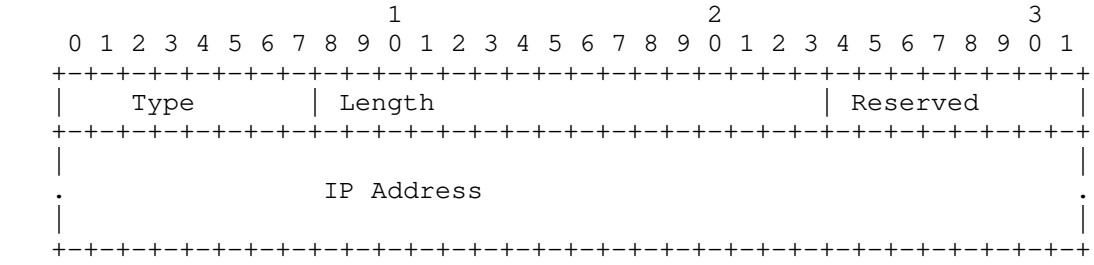


Figure 9 Out of Band IP Address TLV

Type (1 octet) = 64 indicate that this is the TRILL OAM Version

Length (2 octets) = 5 or 17. Length Value 5 indicates it is IPv4 address and Length value of 17 indicates that it is IPv6 address.

IP Address (4 or 16 octets), valid IP address.

4.4.3.1. Diagnostics VLAN TLV

Diagnostic VLAN specifies the VLAN in which the OAM messages are generated. Receiving RBridge MUST compare the inner.VLAN of the Flow entropy to the VLAN specified in the Diagnostic VLAN TLV. Cross connect Flag in the response MUST be set when the two VLANs do not match.

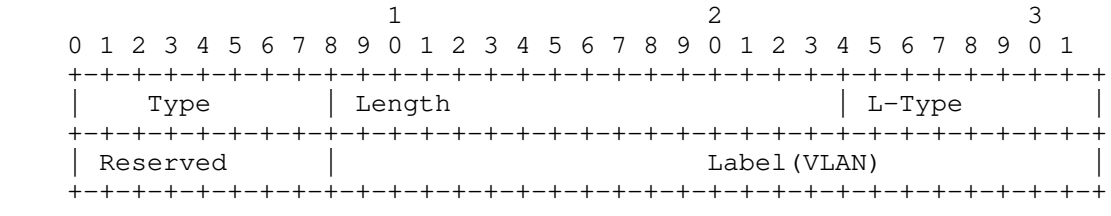


Figure 10 Diagnostic VLAN TLV

Type (1 octet) = 65 indicates that this is the TRILL Diagnostic VLAN TLV

Length (2 octets) = 5

L-Type (Label type, 1 octet)

0- indicate 802.1Q 12 bit VLAN.

1 - indicate TRILL 24 bit fine grain label

Label (24 bits): Either 12 bit VLAN or 24 bit fine grain label.

NOTE: TRILL Operate above the MAC Layer of IEEE 802.1 architecture. Hence it is safe to assume there is no VLAN translation functionality on the inner payload by intermediate RBridges.

4.4.3.2. Original Data Payload TLV

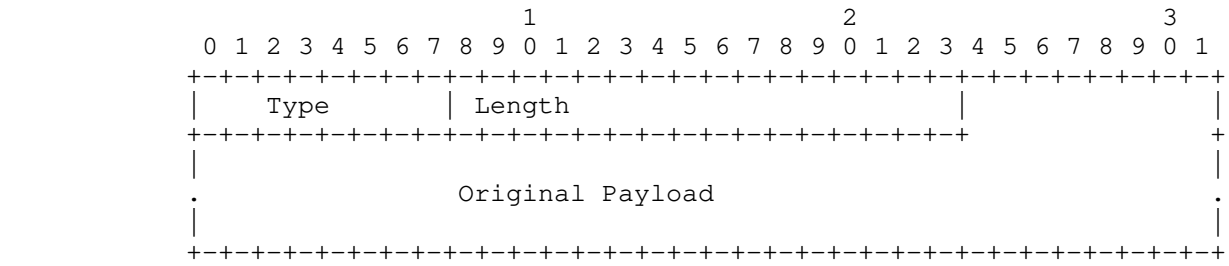


Figure 11 Out of Band IP Address TLV

Length (2 octets) = variable

4.4.3.3. RBridge scope TLV

RBridge scope TLV identifies nicknames of RBridges from which a response is required. RBridge scope TLV only applicable to Multicast Tree Verification messages. This TLV SHOULD NOT be included in other messages. Receiving RBridges MUST ignore this TLV on messages other than Multicast Verification Message.

Each TLV can contain up to 255 nicknames of in scope RBridges. A Multicast Verification Message may contain multiples of "RBridge scope TLVs", in the event more than 255 in scope RBridges needed to be specified.

Absence of the "RBridge scope TLV" indicates, response is needed from all the RBridges. Please see section 7. for details.

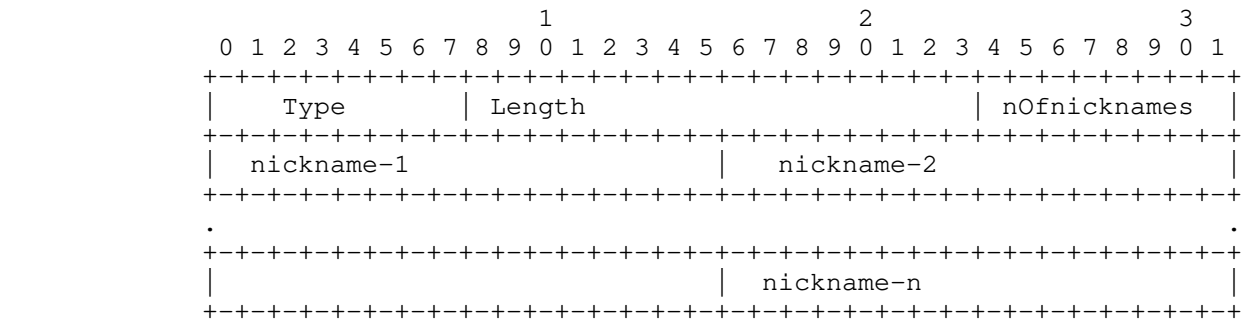


Figure 12 RBridge Scope TLV

Type (1 octet) = 67 indicates that this is the "RBridge scope TLV"

Length (2 octets) = variable. Minimum value is 2.

Nickname (2 octets) = 16 bit RBridge nickname.

4.4.3.4. Upstream RBridge nickname TLV

"Upstream RBridge nickname TLV" identifies nickname or nicknames of the upstream RBridge. [RFC6325] allow a given RBridge to announce multiple nicknames.

"Upstream RBridge nickname TLV" is an optional TLV. Multiple of this TLV MAY be included when an upstream RBridge is represeneted by more than 255 nicknames.

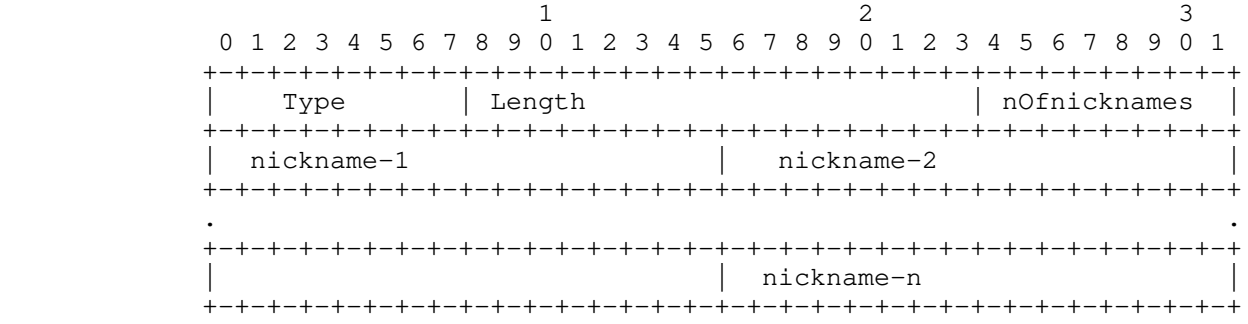


Figure 13 Upstream RBridge nickname TLV

Type (1 octet) = 69 indicates that this is the "Upstream RBridge nickname"

Length (2 octets) = variable. Minimum value is 2.

Nickname (2 octets) = 16 bit RBridge nickname.

4.4.3.5. Next Hop RBridge List TLV

"Next Hop RBridge List TLV" identifies nickname or nicknames of the downstream next hop RBridges. [RFC6325] allows a given RBridge to have mulriple Equal Cost Multi paths to a specified destination.

"Next Hop RBridge List TLV" is an optional TLV. Multiple of this TLV MAY be included when there are more than 255 Equal Cost Paths to the destination.

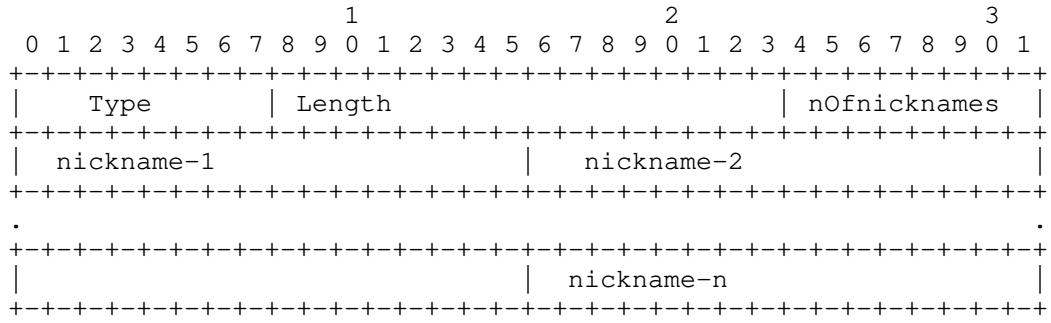


Figure 14 Next Hop RBridge List TLV

Type (1 octet) = 69 indicates that this is the "Next nickname"

Length (2 octets) = variable. Minimum value is 2.

Nickname (2 octets) = 16 bit RBridge nickname.

4.4.3.6. Multicast Receiver Availability TLV

"Multicast Receiver Availability TLV" identifies the number of available multicast receivers available on the responding RBridge on the VLAN specified by the Diagnostic VLAN TLV.

Multicast Receiver Availability is an Optional TLV.

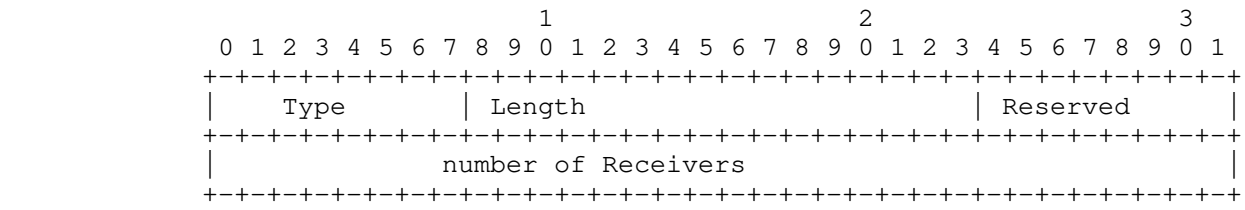


Figure 15 Multicast Receiver Availability TLV

Type (1 octet) = 71 indicates that this is the "Multicast Availability TLV"

Length (2 octets) = 5.

Number if Receivers (4 octets) = Indicates number of Multicast receivers available on the responding RBridge on the VLAN specified by the diagnostic VLAN.

5. Loopback Message

Loopback message is utilized for fault verification. It verifies connectivity between two RBridges, for a specified flow. Additionally, Loopback Message may be utilized for connectivity monitoring and proactive fault detection.

5.1.1. Loopback OAM Message format

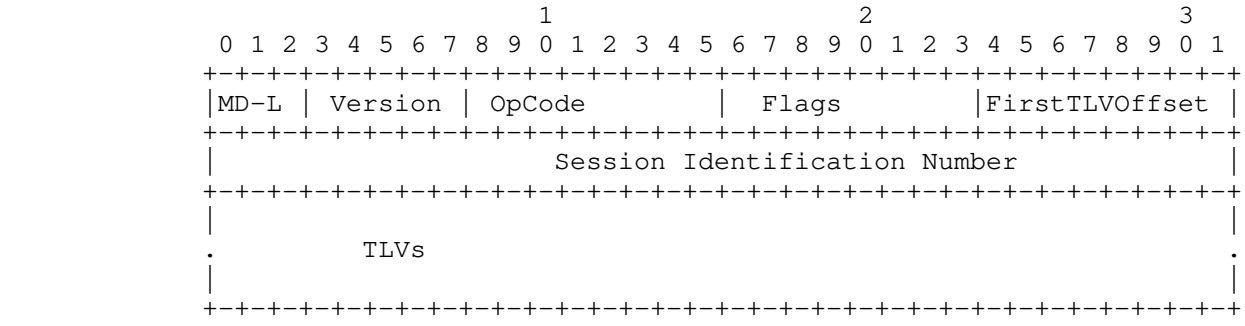


Figure 16 Loopback OAM Message Format

The above figure depicts the format of the Loopback Request and response messages. Opcode for Loopback Message is set to 65 and Opcode of Reply Message is set to 64. Session Identification Number is a 32 bit integer that allows the requesting RBridge to uniquely

identify the corresponding session. Responding RBridges, MUST echo the received "Session Identification" number without modification.

5.1.2. Theory of Operation

5.1.2.1. Originator RBridge

Originator RBridge Identifies the destination RBridge nickname based on user specification or based on location of the specified destination inner MAC address.

Construct the flow entropy based on user specified parameters or implementation specific default parameters.

Construct the TRILL OAM header: Set the opcode to Loopback message type (65). Assign applicable Session Identification number for the request.

TRILL OAM Version TLV MUST be included and set the flags to applicable values.

Include following OAM TLVs, where applicable

- Out-of-band IP address TLV
- Diagnostic VLAN TLV
- Sender ID TLV

Specify the Hop count of the TRILL data frame per user specification or utilize an applicable Hop count value.

Dispatch the OAM frame to the TRILL data plane for transmission.

RBridge may continue to retransmit the request at periodic intervals, until a response is received or the re-transmission count expires. At each transmission Session Identification number MUST be incremented.

5.1.2.2. Intermediate RBridge

Intermediate RBridges forward the frame as a normal data frame and no special handling is required.

5.1.2.3. Destination RBridge

If the Loopback message is addressed to the local RBridge and satisfies OAM identification methods specified in sections Error!

Reference source not found.or 3.2. then the RBridge data plane forwards the message to the CPU for further processing.

TRILL OAM application layer further validates the received OAM frame by examining the presence of OAM-Ethertype at the end of the flow entropy. Frames that do not contain OAM-Ethertype at the end of the flow entropy MUST be discarded.

Construction of the TRILL OAM response:

TRILL OAM application encodes the received TRILL header and flow entropy in the Original payload TLV and includes it in the OAM message.

Set the Return Code and Return sub code to applicable values. Update the TRILL OAM opcode to 64 (Loopback Message Reply)

If the VLAN identifier value of the flow entropy differs from the value specified in the diagnostic VLAN, set the Cross connect Flag on TRILL OAM Application Identifier TLV.

Include the sender ID TLV (1)

If in-band response was requested, dispatch the frame to the TRILL data plane with request-originator RBRidge nickname as the egress RBridge nickname.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

6. Path Trace Message

The Path Trace Message has the same format as Loopback Message. Opcode for Path Trace Reply Message is 66 and Request 67.

Primary use of Path Trace Message is fault isolation. It may also be used for plotting path taken from a given RBridge to another RBridge. Operation of Path Trace message is identical to Loopback message except, that it is first transmitted with a TRILL Hop count field value of 1. Sending RBridge expects a Time Expiry Return-Code from the next hop or a successful response. If a Time Expiry Return-code is received as the response, the originator RBridge records the information received from intermediate node that generated the Time Expiry message and resends the message by incrementing the previous Hop count value by 1. This process is continued until, a response is received from the destination RBridge or Path Trace process timeout occur or Hop count reaches a configured maximum value.

6.1.1. Theory of Operation

6.1.1.1. Originator RBridge

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

Construct the flow entropy based on user specified parameters or implementation specific default parameters.

Construct the TRILL OAM header: Set the opcode to Path Trace Request message type (67). Assign applicable Session Identification number for the request. Return-code and sub-code MUST be set to zero.

TRILL OAM Application Identifier TLV MUST be included and set the flags to applicable values.

Include following OAM TLVs, where applicable

- Out-of-band IP address TLV
- Diagnostic VLAN TLV
- Include the Sender ID TLV

Specify the Hop count of the TRILL data frame as 1 for the first request.

Dispatch the OAM frame to the TRILL data plane for transmission.

RBridge may continue to retransmit the request at periodic interval, until a response received or re-transmission count expires. At each new re-transmission Session Identification number MUST be incremented. Additionally for responses received from intermediate RBridges, RBridge nickname and interface information may be recorded.

6.1.1.2. Intermediate RBridge

Intermediate RBridge receive the Path Trace Messages as Hop count expired frame.

TRILL OAM application layer further validates the received OAM frame by examining the presence of OAM-Ethertype at the end of the flow entropy. Frames that do not contain OAM-Ethertype at the end of the flow entropy MUST be discarded.

Construction of the TRILL OAM response:

TRILL OAM application encodes the received TRILL header and flow entropy in the Original payload TLV and include in the OAM message.

Set the Return Code to (2) "Time Expired" and Return sub code to zero (0). Update the TRILL OAM opcode to 66 (Path Trace Message Reply).

If the VLAN identifier value of the flow entropy differs from the value specified in the diagnostic VLAN, set the Cross connect Flag on TRILL OAM Application Identifier TLV.

Include following TLVs

Upstream RBridge nickname TLV (69)

Reply Ingress TLV (5)

Interface Status TLV (4)

TRILL Next Hop RBridge (Repeat for each ECMP) (70)

Sender ID TLV (1)

If VLAN cross connect error detected, set C flag (Cross connect error detected) in the version.

If in-band response was requested, dispatch the frame to the TRILL data plane with request-originator RBRidge nickname as the egress RBridge nickname.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

6.1.1.3. Destination RBridge

Processing is identical to section 5.1.2.3. With the exception that TRILL OAM Opcode is set to Path Trace Reply (66).

7. Multicast Tree Verification (MTV) Message

Multicast Tree Verification messages allow verifying multicast tree integrity and Multicast address pruning. IGMP snooping is widely deployed in Layer 2 networks for restricting the forwarding of multicast traffic to unwanted destinations. This is accomplished by pruning the multicast tree such that for specified (S,G,VLAN) or (*,G,VLAN), only required destinations are included in the outgoing interface list. It is possible due to timing or implementation

defects, inaccurate pruning of multicast tree, may occur. Such events lead to incorrect multicast connectivity. Multicast tree verification and Multicast group verification messages are design to detect such multicast connectivity defects. Additionally, these tools can be used for plotting a given multicast tree within the TRILL campus.

Multicast tree verification OAM frames are copied to the CPU of every intermediate RBridge that are part of the Multicast tree being verified. Originator of the Multicast Tree verification message, specify the scope of RBridges that a response is required. Only, the RBridges listed in the scope field respond to the request. Other RBridges silently discard the request. Definition of scope parameter is required to prevent receiving large number of responses. Typical scenario of multicast tree verification or group verification involves verifying multicast connectivity to selected set of end-nodes as opposed to the entire network. Availability of the scope, facilitate narrowing down the focus only to the interested RBridges.

Implementations MAY choose to rate limit CPU bound multicast traffic. As result of rate limiting or due to other congestion conditions, MTV messages may be discarded from time to time by the intermediate RBRidges and requester may be required to retransmit the request. Implementations SHOULD narrow the embedded scope of retransmission request only to RBRidges that has failed to respond.

7.1. Multicast Tree Verification (MTV) OAM Message Format

Format of MTV OAM Message format is identical to that of Loopback Message format defined in section 5.1.1.

7.2. Theory of Operation

7.2.1. Originator RBridge

User is required at minimum to specify either the multicast trees that needed to be verified or Multicast MAC address and VLAN or VLAN and Multicast destination IP address. Alternatively, for more specific multicast flow verification, user MAY specify more information e.g. source MAC address, VLAN, Destination and Source IP addresses. Implementation, at minimum, must allow user to specify, choice of multicast trees, Destination Multicast MAC address and VLAN that needed to be verified. Although, it is not mandatory, it is highly desired to provide option to specify the scope. It should be noted source MAC address and some other parameters may not be specified if the Backwards Compatibility Method of section 3.2 is used to identify the OAM frames.

Default parameters MUST be used for unspecified parameters. Flow entropy is constructed based on user specified parameters and/or default parameters.

Based on user specified parameters, originating RBridge identify the nickname that represent the multicast tree.

Obtain the applicable Hop count value for the selected multicast tree.

Construct TRILL OAM message header and include Session Identification number. Session Identification number facilitate the originator to map the response to the correct request.

TRILL OAM Application Identifier TLV MUST be included.

Op-Code MUST be specified as Multicast Tree Verification Message (70)

Include RBridge scope TLV (67)

Optionally, include following TLV, where applicable

- Out-of-band IP address
- Diagnostic VLAN
- Sender ID TLV (1)

Specify the Hop count of the TRILL data frame per user specification. Or utilize the applicable Hop count value, if TRILL Hop count is not being specified by the user.

Dispatch the OAM frame to the TRILL data plane for transmission.

RBridge may continue to retransmit, the request at a periodic interval, until a response received or re-transmission count expires. At each new re-transmission Session Identification number MUS be incremented. At each re-transmission, RBridge may further reduce the scope to the RBridges it has not received a response.

7.2.2. Receiving RBridge

Receiving RBridges identify multicast verification frames per the procedure explained in either section Error! Reference source not found.or section 3.2.

CPU of the RBridge validates the frame and analyzes the scope RBridge list. If the RBrdige scope TLV is present and the local

RBridge nickname is not specified in the scope list, it will silently discard the frame. If the local RBridge is specified in the scope list OR RBridge scope TLV is absent the receiving RBridge proceed for further processing as defined in section 7.2.3.

7.2.3. In scope RBridges

Construction of the TRILL OAM response:

TRILL OAM application encodes the received TRILL header and flow entropy in the Original payload TLV and include in the OAM message.

Set the Return Code to (0) and Return sub code to zero (0). Update the TRILL OAM opcode to 69 (Multicast Tree Verification Reply).

Include following TLVs

Upstream RBridge nickname TLV (69)

Reply Ingress TLV (5)

Interface Status TLV (4)

TRILL Next Hop RBridge (Repeat for each downstream RBridge) (70)

Sender ID TLV (1)

Multicast Receiver Availability TLV (71)

If VLAN cross connect error detected, set C flag (Cross connect error detected) in the version.

If in-band response was requested, dispatch the frame to the TRILL data plane with request-originator RBRidge nickname as the egress RBridge nickname.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

8. Notification Messages

TRILL OAM Notification message format is depicted in following figure.

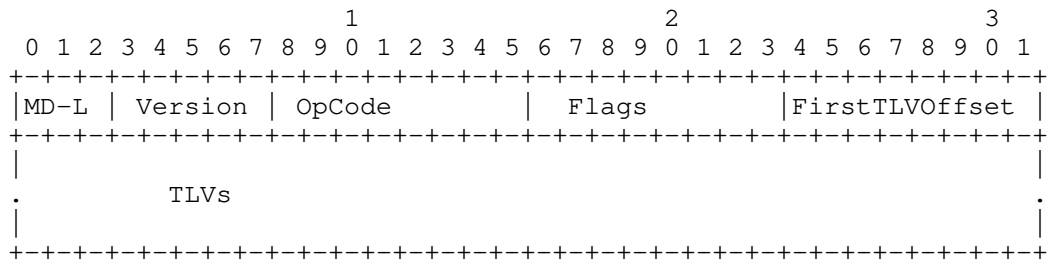


Figure 17 Notification OAM Message Format

The opcode of the Notification message is 68. Notification messages may be generated for variety of errors, warning and informational purposes. Notification messages are almost always asynchronous. Hence there is no Session Identification.

TRILL OAM Application Identifier TLV, which is mandatory, MUST be the first TLV. Return Code and Return sub-code in TRILL OAM version TLV MSUT be set to appropriate value.

9. Return Codes

9.1. Return Codes

Following Return codes are currently defined. These return codes are the initial content of registry setup by IANA. Future allocations are administered by IANA.

0: Success
1: Egress RBridge Nickname unknown
2: Time Expired
3: VLAN Unknown
4: Parameter Problem
5-255: Reserved

9.2. Return sub-codes

For all of the above Return codes, sub-code zero (0) indicates no Return-sub code included.

Currently all other values are reserved and MUST NOT be included unless otherwise specified by IETF publication and registered in IANA.

10. Security Considerations

For general TRILL related security considerations, please refer to [RFC6325]. Specific security considerations related methods presented in this document are currently under investigation.

11. Allocation Considerations

10.1 IEEE Allocation Considerations

The IEEE 802.1 Working Group is requested to allocate a separate opcode and TLV space within 802.1g CFM messages for TRILL purpose.

10.2 IANA Considerations

- Set up sub-registry within the TRILL Parameters registry for block of TRILL OAM OpCodes -

- Set up sub-registry within the TRILL Parameters registry for TRILL OAM TLV Types -

- Set up sub-registry within the TRILL Parameters registry for TRILL OAM return code and return sub codes -

12. References

12.1. Normative References

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

[8021ag] IEEE, "Virtual Bridged Local Area Networks Amendment 5: Connectivity Fault Management", 802.1ag, 2007.

[8021Q] IEEE, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, August st 31 , 2011.

[TRILLOAMFM] Salam, S., et.al., "TRILL OAM Framework", draft-salam-trill-oam-framework-02, Work in Progress, September, 2012.

12.2. Informative References

[RFC6325] Perlman, R., et.al., "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[TRILLML] Senevirathne, T., et.al., "Default Nickname Based Approach for Multi-level TRILL", draft-tissa-trill-multilevel-00, Work in Progress, February 2012.

[RFC6291] Andersson, L., et.al., "Guidelines for the use of the "OAM" Acronym in the IETF" RFC 6291, June 2011.

13. Acknowledgments

Work in this document was largely inspired by the directions provided by Stewart Bryant in finding common OAM solution between SDO.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Tissa Senevirathne
CISCO Systems
375 East Tasman Drive.
San Jose, CA 95134
USA.

Phone: +1 408-853-2291
Email: tsenevir@cisco.com

Samer Salam
CISCO Systems
595 Burrard St. Suite 2123
Vancouver, BC V7X 1J1, Canada

Email: ssalam@cisco.com

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

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

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757

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

Sam Aldrin
Huawei Technologies
2330 Central Express Way
Santa Clara, CA 95951
USA

Email: aldrin.ietf@gmail.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

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

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: April 25, 2013

Mingui Zhang
Huawei
Tissa Senevirathne
CISCO
Janardhanan Pathangi
DELL
Ayan Banerjee
Cumulus Networks
Anoop Ghanwani
DELL

October 22, 2012

TRILL Resilient Distribution Trees
draft-zhang-trill-resilient-trees-01.txt

Abstract

TRILL protocol provides layer 2 multicast data forwarding using IS-IS link state routing. Distribution trees are computed based on the link state information through Shortest Path First calculation and shared among VLANs across the campus. When a link on the distribution tree fails, a campus-wide recovergence of this distribution tree will take place, which can be time consuming and may cause considerable disruption to the ongoing multicast service.

This document proposes to build the backup distribution tree to protect links on the primary distribution tree. Since the backup distribution tree is built up ahead of the link failure, when a link on the primary distribution tree fails, the pre-installed backup forwarding table will be utilized to deliver multicast packets without waiting for the campus-wide recovergence, which minimizes the service disruption.

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

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction 4
1.1. Conventions used in this document 5
1.2. Terminology 5
2. Usage of Affinity TLV 5
2.1. Allocating Affinity Links 5
2.2. Distribution Tree Calculation with Affinity Links 6
3. Resilient Distribution Trees Calculation 7
3.1. Designating Roots for Backup Trees 8
3.1.1. Conjugate Trees 8
3.1.2. Explicitly Advertising Tree Roots 8
3.2. Backup DT Calculation 9
3.2.1. Backup DT Calculation with Affinity Links 9
3.2.1.1. Algorithm for Choosing Affinity Links 9
3.2.1.2. Affinity Links Advertisement 10
3.2.2. Backup DT Calculation without Affinity Links 10
4. Resilient Distribution Trees Installation 10
4.1. Pruning the Backup Distribution Tree 11
4.2. RPF Filters Preparation 11
5. Protection Mechanisms with Resilient Distribution Trees 12
5.1. Global 1:1 Protection 13
5.2. Global 1+1 Protection 13
5.2.1. Failure Detection 13
5.2.2. Traffic Forking and Merging 14
5.3. Local Protection 14

5.3.1. Start Using Backup Distribution Tree 15
5.3.2. Duplication Suppression 15
5.3.3. An Example to Walk Through 15
5.4. Switching Back to the Primary Distribution Tree 16
6. Security Considerations 17
7. IANA Considerations 17
8. References 17
8.1. Normative References 17
8.2. Informative References 18
Author's Addresses 19

1. Introduction

Lots of multicast traffic is generated by interrupt latency sensitive applications, e.g., video distribution, including IP-TV, video conference and so on. Normally, network fault will be recovered through a network wide reconvergence of the forwarding states but this process is too slow to meet the tight SLA requirements on the service disruption duration. What is worse, updating multicast forwarding states may take significantly longer than unicast convergence since multicast states are updated based on control-plane signaling [mMRT].

Protection mechanisms are commonly used to reduce the service disruption caused by network fault. With backup forwarding states installed in advance, a protection mechanism is possible to restore a interrupted multicast stream in tens of milliseconds which guarantees the stringent SLA on service disruption. Several protection mechanisms for multicast traffic have been developed for IP/MPLS networks [mMRT] [MoFRR]. However, the way TRILL constructs distribution trees (DT) is different from the way multicast trees are computed under IP/MPLS therefore a multicast protection mechanism suitable for TRILL is required.

This document proposes "Resilient Distribution Trees (RDT)" in which backup trees are installed in advance for the purpose of fast failure repair. Three types of protection mechanisms are proposed. Global 1:1 protection is used to refer to the mechanism having the multicast source RBridge normally injects one multicast stream onto the primary DT. When this stream is detected to be interrupted, the source RBridge switches to the backup DT to inject subsequent multicast stream until the primary DT is recovered. Global 1+1 protection is used to refer to the mechanism having the multicast source RBridge always injects two copies of multicast streams onto the primary DT and backup DT respectively. In normal case, multicast receivers pick the stream sent along the primary DT and egress it to its local link. When a link failure interrupts the primary stream, the backup one will be picked until the primary DT is recovered. Local protection refers to the mechanism having the RBridge attached to the failed link to locally repair the failure.

RDT may greatly reduce the service disruption caused by link failures. In the global 1:1 protection, the time cost by DT recalculation and installation can be saved. The global 1+1 protection and local protection further save the time spent on failure propagation. A failed link can be repaired in tens of milliseconds. Although it’s possible to make use of RDT to achieve load balance of multicast traffic, this document leaves it behind for future study.

[6326bis] defines the Affinity TLV. An "Affinity Link" can be explicitly assigned to a distribution tree or trees. This offers a way to manipulate the calculation of distribution trees. With intentional assignment of Affinity Links, a backup distribution tree can be set up to protect links on a primary distribution tree.

1.1. 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].

1.2. Terminology

IS-IS: Intermediate System to Intermediate System

TRILL: TRansparent Interconnection of Lots of Links

DT: Distribution Tree

RPF: Reverse Path Forwarding

RDT: Resilient Distribution Tree

SPF: Shortest Path First

SPT: Shortest Path Tree

PRB: the Parent RBridge attached to a link on a distribution tree

PLR: Point of Local Repair, in this document, it is the multicast upstream RBridge connecting the failed link. It’s valid only for local protection.

2. Usage of Affinity TLV

The Affinity TLV is currently only used to assign parents for leaf nodes [6326bis]. This document expands the scope of its usage to assign a parent to a non-leaf RBridge without changing the definition of this TLV.

2.1. Allocating Affinity Links

Affinity TLV explicitly assigns parents for RBridges on distribution trees. They are advertised in the Affinity TLV and recognized by each RBridge in the campus. The originating RBridge becomes the parent and the nickname contained in the Affinity Record identifies the child, which explicitly provides an "Affinity Link" on a distribution tree

or trees. The "Tree-num of roots" of the Affinity Record identify the distribution trees that adopt this Affinity Link [6326bis].

Affinity Links may be configured or automatically determined using a certain algorithm [CMT]. Suppose link RB2-RB3 is chosen as an Affinity Link on the distribution tree rooted at RB1. RB2 should send out the Affinity TLV with an Affinity Record like {Nickname=RB3, Num of Trees=1, Tree-num of roots=RB1}. In this document, RB3 does not have to be a leaf node on a distribution tree, therefore an Affinity Link can be used to identify any link on a distribution tree. This kind of assignment offers a flexibility to RBridges in distribution tree calculation: they are allowed to choose parents not on the shortest paths to the root. This flexibility is leveraged to increase the reliability of distribution trees in this document.

2.2. Distribution Tree Calculation with Affinity Links

When RBridges receive an Affinity Link which is an incoming link of RB2. RB2’s incoming links other than the Affinity Link are removed from the full graph of the campus to get a sub graph. RBridges perform Shortest Path First (SPF) calculation to compute the distribution tree based on the sub graph. In this way, the Affinity Link will appear on the distribution tree.

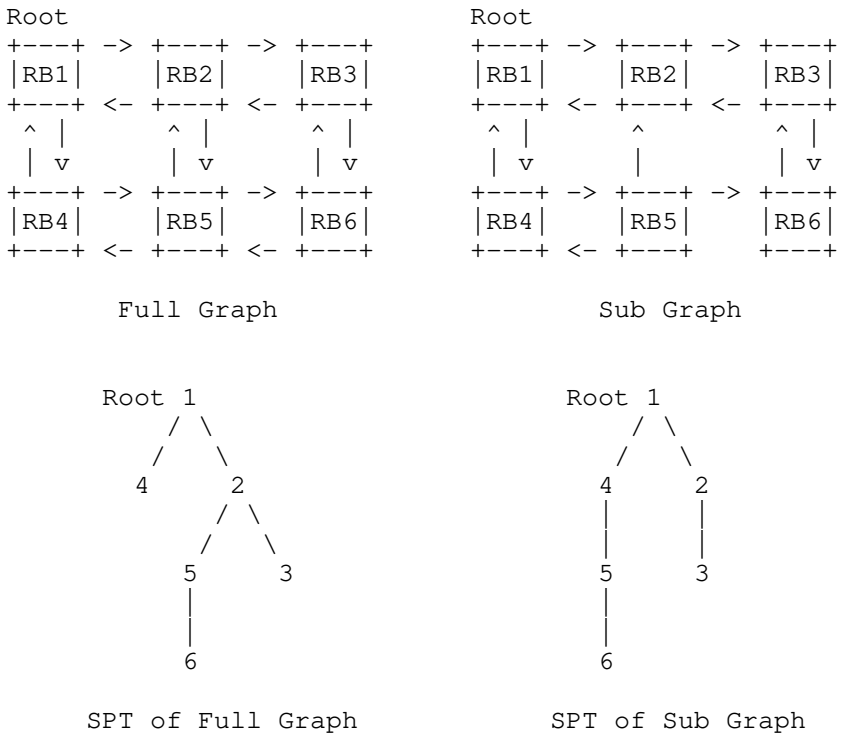


Figure 2.1: DT Calculation with the Affinity Link RB4-RB5

Take Figure 2.1 as an example. Suppose RB1 is the root and link RB4-RB5 is the Affinity Link. RB5's other incoming links RB2-RB5 and RB6-RB5 are removed from the Full Graph to get the Sub Graph. Since RB4-RB5 is the unique link to reach RB5, the Shortest Path Tree (SPT) inevitably contain this link.

3. Resilient Distribution Trees Calculation

RBridges leverage IS-IS to detect and advertise network fault. A node or link failure will trigger a campus-wide reconvergence of distribution trees. The reconvergence generally includes the following procedures:

1. Failure detected through IS-IS control messages (HELLO) exchanging;

2. IS-IS flooding and each RBridge recognizes the failure;

3. Each RBridge recalculates affected distribution trees independently;

4. RPF filters are updated according to the new distribution trees. The recomputed distribution trees are pruned per VLAN and installed into the multicast forwarding tables.

The slow reconvergence can be as long as tens of seconds or even minutes, which will cause disruption to ongoing multicast traffic. In protection mechanisms, alternative paths prepared ahead of potential node or link failures are leveraged to detour the failures upon the failure detection, therefore service disruption can be minimized.

In order to protect a node on the primary tree, a backup tree can be set up as lack of this node [mMRT]. When this node fails, the backup tree can be safely used to forward multicast traffic to make a detour. However, TRILL distribution trees are shared among all VLANs and they have to cover all RBridge nodes in the campus [RFC6325]. A DT does not span all RBridges in the campus may not cover all receivers of many a multicast group (This is different from the multicast trees construction signaled by PIM [RFC4601] or mLDP [RFC6388].). Therefore, the construction of backup DT for the purpose of node protection is out the scope of this document. This document will focus only on link protection from now on.

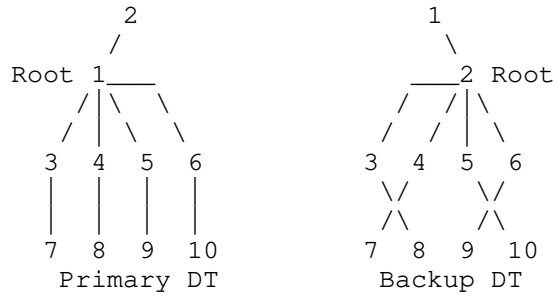


Figure 3.1: An Example of a Primary DT and its Backup DT

3.1. Designating Roots for Backup Trees

Operators MAY manually configure the roots for the backup DTs. Nevertheless, this document aims to provide a mechanism with minimum configuration. Two options are offered as follows.

3.1.1. Conjugate Trees

RFC 6325 has defined how distribution tree roots are selected. When a backup DT is computed for a primary DT, its root is set to be the root of this primary DT.

3.1.2. Explicitly Advertising Tree Roots

RBridge RB1 having the highest root priority nickname might explicitly advertise a list of nicknames to identify the roots of the primary and backup tree roots (See RFC6325 Section 4.5).

3.2. Backup DT Calculation

3.2.1. Backup DT Calculation with Affinity Links

TRILL allows RBridges to compute multiple distribution trees. With the intentional assignment of Affinity Links in DT calculation, this document proposes the method to construct Resilient Distribution Trees (RDT). For example, in Figure 3.1, the backup DT is set up maximally disjoint to the primary DT (The full topology is an combination of these two DTs, which is not shown in the figure.). Except the link between RB1 and RB2, all other links on the primary DT do not overlap with links on the backup DT. It means that every link on the primary DT except link RB1-RB2 can be protected by the backup DT.

3.2.1.1. Algorithm for Choosing Affinity Links

Operators MAY configure Affinity Links to intentionally protect a specific link, such as the link connected to a gateway. But it is desirable that each RBridge independently computes Affinity Links for a backup DT while the same result is got across the whole campus, which enables a distributed deployment and also minimizes configuration .

Algorithms for MRT [mMRT] may be used to figure out Affinity Links on a backup DT which is maximally disjoint to the primary DT but it only provides a subset of all possible solutions. In TRILL, RDT does not restrict that the root of the backup DT is the same as that of the primary DT. Two disjoint (or maximally disjoint) trees may root from different nodes, which significantly augments the solution space.

This document RECOMMENDS to achieve the independent method through a slight change to the conventional DT calculation process of TRILL. Basically, after the primary DT is calculated, the RBridge will be aware of which links will be used. When the backup DT is calculated, each RBridge increases the metric of these links by a proper value (for safety, the summation of all original link metrics in the campus is RECOMMENDED), which gives these links a lower priority being chosen by the backup DT by performing SPF calculation. All links on this backup DT can be assigned as Affinity Links but this is unnecessary. In order to reduce the amount of Affinity TLVs flooded across the campus, only those will not picked by conventional DT calculation process ought to be recognized as Affinity Links.

3.2.1.2. Affinity Links Advertisement

Similar as [CMT], every Parent RBridge (PRB) of an Affinity Link take charge of announcing this link in the Affinity TLV. When this RBridge plays the role of PRB for several Affinity Links, it is natural to have them advertised together in the same Affinity TLV and each Affinity Link is structured as one Affinity Record.

Affinity Links are announced in the Affinity TLV which is recognized by every RBridge. Since each RBridge computes distribution trees as the Affinity TLV requires, the backup DT will built up naturally.

3.2.2. Backup DT Calculation without Affinity Links

This section aims to provide an alternative method to set up the disjoint DT without Affinity Links.

After the primary DT is calculated, each RBridge increases the weights of those links which are already in the primary DT by a multiplier (For safety, 100x is RECOMMENDED.). That would ensure that a link appears in 2 trees if and only if there is no other way to reach the node (i.e. the graph would become disconnected if it were pruned of the links in the first tree.). In other words, the two trees will be maximally disjoint.

The above algorithm is similar as that defined in Section 3.2.1.2. All RBridges MUST agree on this algorithm, then backup distribution trees can be automatically calculated by each RBridge and configuration is unnecessary.

4. Resilient Distribution Trees Installation

As specified in RFC 6325 Section 4.5.2, an ingress RBridge MUST announce the distribution trees it may choose to ingress multicast frames. Thus other RBridges in the campus can limit the amount of states which are necessary for RPF check. Also, RFC 6325 recommends that an ingress RBridge chooses the DT or DTs whose root or roots are least cost from the ingress RBridge. To sum up, RBridges do precompute all the trees that might be used but only install part of them according to each ingress.

This document states that the backup DT MUST be contained in an ingress RBridge’s DT announcing list and included in this ingress RBridge’s LSP. In order to reduce the service disruption time, RBridges SHOULD install backup DTs in advance, which also includes the RPF filters that need to be set up for RPF Check.

Since the backup DT is intentionally built up maximally disjoint to

the primary DT, when a link fails and interrupts the ongoing multicast traffic sent along the primary DT, it is probably that the backup DT is not affected. Therefore, the backup DT installed in advance can be used to deliver multicast frames immediately.

4.1. Pruning the Backup Distribution Tree

Backup DT should be pruned per-VLAN. But the way backup DT being pruned is different from the way that the primary DT is pruned. Even though a branch contains no downstream receivers, it is probably that it should not be pruned for the purpose of protection. Therefore, a branch on the backup DT should be pruned per-VLAN, eliminating branches that have no potential downstream RBridges which appear on the pruned primary DT.

It is probably that the primary DT is not optimally pruned in practice. In this case, the backup DT SHOULD be pruned presuming that the primary DT is optimally pruned. Those redundant links ought to be pruned will not be protected.

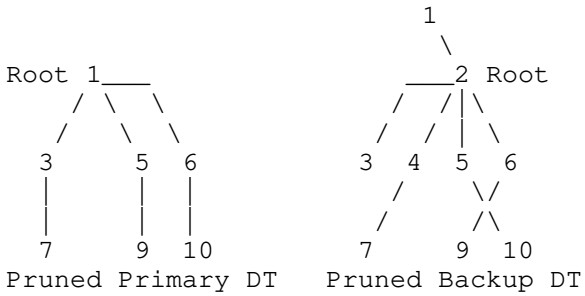


Figure 4.1: The Backup DT is Pruned Based on the Pruned Primary DT.

Suppose RB7, RB9 and RB10 constitute a multicast group. The pruned primary DT and backup DT are shown in Figure 4.1. Branches RB2 and RB4 on the primary DT are pruned since there are no potential receivers on these two branches. Although branches RB1 and RB3 on the backup DT have no potential multicast receivers, they may be used to repair link failures of the primary DT. Therefore they are not pruned from the backup DT. Branch RB8 can be safely pruned because it does not appear on the pruned primary DT.

4.2. RPF Filters Preparation

RB2 includes in its LSP the information to indicate which trees RB2 might choose to ingress multicast frames [RFC6325]. When RB2 specifies the trees it might choose to ingress multicast traffic, it SHOULD include the backup DT. Other RBridges will prepare the RPF

check states for both the primary DT and backup DT. When a multicast packet is sent along either the primary DT or the backup DT, it will pass the RPF Check. This works when global 1:1 protection is used. However, when global 1+1 protection or local protection is applied, traffic duplication will happen if multicast receivers accept both copies of multicast frame from two RPF filters. In order to avoid such duplication, multicast receivers (egress RBridge) MUST act as merge points to active a single RPF filter and discard the duplicated frames from the other RPF filter. In normal case, the RPF state is set up according to the primary DT. When a link fails, the RPF filter should be updated instantly according to the backup DT.

5. Protection Mechanisms with Resilient Distribution Trees

Protection mechanisms can be developed to make use of the backup DT installed in advance. But protection mechanisms already developed using PIM or mLDP for multicast of IP/MPLS networks are not applicable to TRILL due to the following fundamental differences in their distribution tree calculation.

- The link on a TRILL distribution tree is bidirectional while the link on a distribution tree in IP/MPLS networks is unidirectional.

- In TRILL, an multicast source node does not have to be the root of the distribution tree. It goes just the opposite in IP/MPLS networks.

- In IP/MPLS networks, distribution trees are constructed for each multicast source node as well as their backup distribution trees. In TRILL, a small number of core distribution trees are shared among multicast groups. A backup DT does not have to share the same root as the primary DT.

Therefore TRILL needs dedicated multicast protection mechanisms.

Global 1:1 protection, global 1+1 protection and local protection are developed in this section. In Figure 4.1, assume RB7 is the ingress RBridge of the multicast stream while RB9 and RB10 are the multicast receivers. Suppose link RB1-RB5 fails during the multicasting. The backup DT rooted at RB2 does not include the link RB1-RB5, therefore it can be used to protect this link. In the global 1:1 protection, RB7 will switch the subsequent multicast traffic to this backup DT when it's notified about the link failure. In the global 1+1 protection, RB7 will inject two copies of the multicast stream and let multicast receivers RB9 and RB10 merge them. In the local protection, when link RB1-RB5 fails, RB1 will locally replicate the multicast traffic and send it on the backup DT.

5.1. Global 1:1 Protection

In the global 1:1 protection, the ingress of the multicast traffic is responsible to switch the failure affected traffic from the primary DT over to the backup DT. Since the backup DT has been installed in advance, the global protection does not need to wait for the DT recalculation and installation. Upon the ingress RBridge is notified about the failure, it immediately makes this switch over.

This type of protection is simple and duplication safe. However, depending on the topology of the RBridge campus, the time spent on the failure detection and propagation through the IS-IS control plane may still cause considerable service disruption.

BFD (Bidirectional Forwarding Detection) protocol can be used to reduce the failure detection time [rbBFD]. Multi-destination BFD extends BFD mechanism to include the fast failure detection of multicast paths [mBFD]. It can be used to reduce both the failure detection and propagation time in the global protection. In multi-destination BFD, ingress RBridge need to send BFD control packets to poll each receiver, and receivers return BFD control packets to the ingress as response. If no response is received from a specific receiver for a detection time, the ingress can judge that the connectivity to this receiver is broken. In this way, multi-destination BFD detects the connectivity of a path rather than a link. The ingress RBridge will determine a minimum failed branch which contains this receiver. Ingress RBridge will switch ongoing multicast traffic based on this judgment. For example, if RB9 does not response while RB10 still responses, RB7 will presume that link RB1-RB5 and RB5-RB9 are failed. Multicast traffic will be switched to a backup DT that can protect these two links. Accurate link failure detection might help ingress RBridge to make smarter decision but it's out of the scope of this document.

RBridges may make use of RBridge Channel to speed up the failure propagation [RBch]. LSPs for the purpose of failure notification may be sent to the ingress RBridge as unicast TRILL Data using RBridge Channel.

5.2. Global 1+1 Protection

In the global 1+1 protection, the multicast source RBridge always replicate the multicast frames and send them onto both the primary and backup DT. This may sacrifice the capacity efficiency but given there is much connection redundancy and inexpensive bandwidth in Data Center Networks, such kind of protection can be popular [MoFRR].

5.2.1. Failure Detection

Egress RBridges (merge points) SHOULD realize the link failure as early as possible so that failure affected egress RBridges may update their RPF filters quickly to minimize the traffic disruption. Three options are provided as follows.

1. Egress RBridges assume a minimum known packet rate for a given data stream [MoFRR]. A failure detection timer Td are set as the interval between two continuous packets. Td is reinitialized each time a packet is received. If Td expires and packets are arriving at the egress RBridge on the backup DT (within the time frame Td), it updates the RPF filters and starts to receive packets forwarded on the backup DT.

2. With multi-destination BFD, when a link failure happens, affected egress RBridges can detect a lack of connectivity from the ingress [mBFD]. Therefore these egress RBridges are able to update their RPF filters promptly.

3. Egress RBridges can always rely on the IS-IS control plane to learn the failure and determine whether their RPF filters should be updated.

5.2.2. Traffic Forking and Merging

For the sake of protection, transit RBridges SHOULD active both primary and backup RPF filters, therefore both copies of the multicast frames will pass through transit RBridges.

Multicast receivers (egress RBridges) MUST act as "merge points" to egress only one copy of these multicast frames. This is achieved by the activation of only a single RPF filter. In normal case, egress RBridges will activate the primary RPF filter. When a link on the pruned primary DT fails, ingress RBridge cannot reach some of the receivers. When these unreachable receivers realize it, they SHOULD update their RPF filters to receive packets sent on the backup DT.

5.3. Local Protection

In the local protection, the Point of Local Repair (PLR) happens at the upstream RBridge connecting the failed link who makes the decision to replicate the multicast traffic to recover this link failure. Local protection can further save the time spent on failure notification through the flooding of LSPs across the campus. In addition, the failure detection can be speeded up using BFD [RFC5880], therefore local protection can minimize the service disruption within 50 milliseconds.

Since the ingress RBridge is not necessarily the root of the

distribution tree in TRILL, a multicast downstream point may be not the descendants of the ingress point on the distribution tree. Moreover, distribution trees in TRILL are bidirectional and do not share the same root. There are fundamental differences between the distribution tree calculation of TRILL and those used in PIM and mLDP, therefore local protection mechanisms used for PIM and mLDP, such as [mMRT] and [MoFRR], are not applicable to TRILL.

5.3.1. Start Using Backup Distribution Tree

The egress nickname of the replicated multicast TRILL data frames will be rewritten to the backup DT's root nickname by the PLR. But the ingress of the multicast frame MUST be remained unchanged. This is a halfway change of the DT for multicast frames. Then the PLR begins to forward multicast traffic along the backup DT (same ingress but different egress).

In the above example, if PLR RB1 decides to send replicated multicast frames according to the backup DT, it will send it to the next hop RB2. However, according to the RPF filter built up from the backup DT, multicast frames ingressed by RB7 should only be received from the link RB4-RB2. So RB2 will discard these frames. In fact, any RBridge should receive multicast frames from any ingress, through a single link. The halfway change of DT must modify this rule in order to be valid. When RB20 computes the RPF filter for each ingress RB30 for the backup DT, RB20 believes any link on the backup DT connecting RB20 may be the link on which RB20 may receive a packet from RB30. In this way, in the above example RB2 will not discard the multicast frames sent from RB1.

5.3.2. Duplication Suppression

When a PLR starts to sent replicated multicast frames on the backup DT, multicast frames sent along the primary DT are still going on. Some RBridges on the primary DT might receive two copies of these multicast frames, filled with two different egress nicknames. Local protection MUST adopt duplication suppression mechanism such as the traffic forking and merging method in the global 1+1 protection.

5.3.3. An Example to Walk Through

The example used in the above local protection is put together to get a whole "walk through" below.

In the normal case, multicast frames ingressed by RB7 using the pruned primary DT rooted at RB1 are being received by RB9 and RB10. When the link RB1-RB5 fails, the PLR RB1 begins to replicate and forward subsequent multicast frames using the pruned backup DT rooted

at RB2. When RB2 gets the multicast frames from the link RB1-RB2, it accepts them since the RPF filter {DT=RB2, ingress=RB7, receiving links=RB1-RB2, RB3-RB2, RB4-RB2, RB5-RB2 and RB6-RB2} is installed on RB2. RB2 forwards the replicated multicast frames to its neighbors except RB1. When the multicast frames reach RB6 where both RPF filters {DT=RB1, ingress=RB7, receiving link=RB1-RB6} and {DT=RB2, ingress=RB7, receiving links=RB2-RB6 and RB9-RB6} are active. RB6 will let both multicast streams through. Multicast frames will finally reach RB9 where the RPF filter is updated from {DT=RB1, ingress=RB7, receiving link=RB5-RB9} to {DT=RB2, ingress=RB7, receiving link=RB6-RB9}. RB9 will egress the multicast frames on to the local link.

From the above explanation, we can find that we have to change the data plane with egress rewriting and relax the RPF Checking for the local protection.

5.4. Switching Back to the Primary Distribution Tree

Assume an RBridge receives the LSP which indicates the link failure. This RBridge starts to calculate the new primary DT based on the topology with the failed link. Suppose the new primary DT is installed at t1.

The propagation of LSPs around the campus takes time. For safety, we assume all RBridges in the campus have converged to the new primary DT at t1+Ts (By default, Ts is set to 30s.). At t1+Ts, the ingress RBridge switches the traffic from the backup DT back to the new primary DT.

After another Ts (at t1+2*Ts), no multicast frames are being forwarded along the old primary DT. The backup DT SHOULD be updated according to the new primary DT. The process of this update under different protection types are discussed as follows.

a) For the global 1:1 protection, the backup DT is simply updated at t1+2*Ts.

b) For the global 1+1 protection, the ingress RBridge has stopped replicating the multicast frames onto the old backup DT at t1+Ts. The backup DT is updated at t1+2*Ts. It MUST wait for another Ts, during which time period all RBridges converge to the new backup DT. At t1+3*Ts, the ingress RBridge MAY start to replicate multicast frame onto the new backup DT.

c) For the local protection, the PLR may stop replicating and sending packets on the old backup DT at t1+Ts. However, if the PLR stops redirecting earlier than the ingress RBridge switches to the new

primary DT, packet loss may happen; If the PLR stops too late, frame duplication may happen. In a special case as mentioned in [mMRT], the destination end-station is able to resolve the frame duplication. Then the PLR may stop the redirecting at t1+2*Ts. After t1+3*Ts, RBridges may begin to update the backup DT.

6. Security Considerations

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

7. IANA Considerations

No new registry is requested to be assigned by IANA. The Affinity TLV has already been defined in [6326bis]. This document does not change its definition. RFC Editor: please remove this section before publication.

8. References

8.1. Normative References

[6326bis] D. Eastlake, A. Banerjee, et al., "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-07.txt, work in Progress.

[CMT] T. Senevirathne, J. Pathangi, et al, "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-trill-cmt-00.txt, working in progress.

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

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

[rbBFD] V. Manral, D. Eastlake, et al, "TRILL (Transparent Interconnetion of Lots of Links): Bidirectional Forwarding Detection (BFD) Support", draft-ietf-trill-rbridge-bfd-06.txt, work in progress.

[mBFD] D. Katz, D. Ward, "BFD for Multipoint Networks", draft-

ietf-bfd-multipoint-00.txt, work in progress.

[RFC5880] D. Katz, D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

8.2. Informative References

[mMRT] A. Atlas, R. Kebler, et al., "An Architecture for Multicast Protection Using Maximally Redundant Trees", draft-atlas-rtgwg-mrt-mc-arch-00.txt, work in progress.

[MoFRR] A. Karan, C. Filsfils, et al., "Multicast only Fast Re-Route", draft-karan-mofrr-02.txt, work in progress.

[RBch] D. Eastlake, V. Manral, et al, "TRILL: RBridge Channel Support", draft-ietf-trill-rbridge-channel-06.txt, work in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies Co.,Ltd
Huawei Building, No.156 Beiqing Rd.
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Tissa Senevirathne
Cisco Systems
375 East Tasman Drive,
San Jose, CA 95134

Phone: +1-408-853-2291
Email: tsenevir@cisco.com

Janardhanan Pathangi
Dell/Force10 Networks
Olympia Technology Park,
Guindy Chennai 600 032

Phone: +91 44 4220 8400
Email: Pathangi_Janardhanan@Dell.com

Ayan Banerjee
Cumulus Networks
1089 West Evelyn Avenue
Sunnyvale, CA 94086 USA

EMail: ayabaner@gmail.com

Anoop Ghanwani
Dell
350 Holger Way
San Jose, CA 95134

Phone: +1-408-571-3500
Email: Anoop@alumni.duke.edu