

TRILL working group  
Internet Draft  
Intended status: Standard Track  
Expires: Sept 2012

L. Dunbar  
D. Eastlake  
Huawei  
Radia Perlman  
Intel  
I. Gashinsky  
Yahoo  
July 11, 2011

Directory Assisted RBridge edge  
draft-dunbar-trill-directory-assisted-edge-01.txt

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/lid-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 January 11, 2009.

Copyright Notice

Copyright (c) 2011 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 BSD License.

## Abstract

RBridge edge nodes currently learn the mapping between MAC address and its corresponding RBridge edge node address by observing the data packets traversed through.

This document describes why and how directory assisted RBridge edge nodes can improve TRILL network scalability in data center environment.

## 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 0.

## Table of Contents

1. Introduction .....	3
2. Terminology .....	3
3. Impact to RBridge Network by massive number of hosts in Data Center(s) .....	3
4. Directory Assisted RBridge Edge in Data Center.....	5
5. Further optimization in using directory assistance.....	7
5.1. TRILL Header encapsulated by non-RBridge nodes.....	9
6. Conclusion and Recommendation.....	10
7. Manageability Considerations.....	10
8. Security Considerations.....	10
9. IANA Considerations .....	10
10. Acknowledgments .....	11
11. References .....	11
Authors' Addresses .....	11
Intellectual Property Statement.....	12
Disclaimer of Validity .....	13

## 1. Introduction

Data center networks are different from campus networks in several ways. Main differences include:

- o Data centers, especially Internet or cloud data centers with virtualized servers, tend to have large number of hosts
- o Topology is based on racks, rows.
  - Hosts assignment to Servers, Racks, and Rows is orchestrated by Server/VM Management system, not random.
- o With virtualization, there is an ever increasing trend to dynamically create VMs when the application requires more resources, and move VMs, either from overloaded servers, or to aggregate VMs onto fewer servers to save power when demand is light. This may lead to hosts belonging to same subnet being placed under different locations (racks or rows).

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

## 2. Terminology

AF        Appointed Forwarder RBridge port

Bridge:   IEEE802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

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

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

VM:       Virtual Machines

## 3. Impact to RBridge Network by massive number of hosts in Data Center(s)

In a data center, there are likely to be very large number of hosts (e.g. hundreds of thousands, or even more) and hosts belonging to one subnet may be placed under different racks or rows. If TRILL is deployed in those data centers, hosts belonging to one subnet may be

placed under multiple edge RBridges which are on different Bridged LAN. And each edge RBridge needs to enable multiple VLANs. This creates several problems which this draft is intending to address:

- Unnecessary filling of slots in MAC table of edge RBridges, due to an edge RBridge, R1 receiving broadcast traffic (ARP/ND queries or gratuitous APRs) from hosts that are not actually communicating with any hosts attached to R1.
- The current TRILL protocol requires a MAC address to only be accessible from one RBridge edge port (AF port). When a data center has dual uplinks for each rack of servers to two different Access switches (which is very common), some links can't be fully utilized.
- Flooding within RBridge domain triggered by ARP/ND.

Consider a data center with 80 rows, 8 racks per row and 40 servers per rack. There can be  $80 \times 8 \times 40 = 25600$  servers. Suppose each server is virtualized to 20 VMs, there could be  $25600 \times 20 = 512000$  hosts in this data center. A common network design for this kind of data centers is to have multiple tiers of switches, e.g. one or two Access Switches for each rack (ToR), Aggregation switches for each row (or EoR), and some Core switches to interconnect the Aggregation switches.

If TRILL is to be deployed in this data center, let's consider following two scenarios of TRILL domain boundary:

- Scenario #1: TRILL domain boundary are Access (TOR) switches, i.e. ToR being the edge RBridges:

With 80 rows and 8 racks per row, there will be  $80 \times 8 = 640$  edge RBridges, with each Edge RBridge supporting 40 RBridge edge ports (facing the servers) and 8 RBridge trunk ports facing aggregation (EoR) switches. Then there are  $40 \times 640 = 25600$  RBridge edge ports in this data center.

If each rack and row has two redundant switches, then there will be  $640 \times 2 = 1280$  RBridge edge nodes and  $80 \times 2 = 160$  RBridge core nodes. Total number of nodes in this RBridge domain could be 1440 ( $1280 + 160$ ) plus some core switches which interconnect all the aggregation (EoR) switches very large number of nodes in this RBridge IS/IS domain.

- Scenario #2: TRILL domain boundary are the aggregation (EoR) switches:

With the same assumption as before, there will be 80 Edge RBridges in the RBridge IS/IS domain. Even with redundancy, the number of nodes in RBridge domain will be less than 200. Therefore, the size of the RBridge IS/IS domain is reasonable.

But, this scenario creates a Bridged LAN attached to RBridge edge ports. It becomes necessary to designate only one port (AF port) to forward native traffic to avoid loops among multiple RBridge edge ports and to have some mechanisms to prevent loops within the Bridged LAN attached to RBridge edge ports. Designating one AF port for forwarding native traffic not only makes some links unusable but also put extra heavy load on the AF port. In addition, when AF changes, traffic temporarily goes to black hole. Running traditional Layer 2 STP/RSTP on the Bridged LAN for loop prevention may be overkill because the topology among the ToR switches and RBridge edge is very simple. This draft proposes a simple directory assisted approach to avoid loops.

In addition, the number of MAC&VLAN<->RBridge Edge Mappings to be learned and managed by RBridge edge node can be very large. In the example above, each Edge RBridge has 8 RBridge edge ports facing the ToR switches. Since each ToR has 40 downstream ports facing servers and each server has 20 VMs, there are  $40 \times 20 = 800$  hosts attached to each downstream port of an EoR switch and total of  $8 \times 800 = 6400$  hosts attached to this EoR switch. If all those 6400 hosts belong to 640 VLANs and each VLAN has 200 hosts, then, under the worst case scenario, the total number of MAC&VLAN entries to be learned by the RBridge edge (i.e. EoR) can be  $640 \times 200 = 128000$ . You can easily see that the number of MAC&VLAN<->RBridge Edge mapping entries to be learnt by the RBridge edge node can be very large.

#### 4. Directory Assisted RBridge Edge in Data Center environment

In data center environment, the hosts (VMs) placement to servers, racks, and rows is orchestrated by Server (or VM) Management System(s), i.e. there is a database or multiple ones (distributed model) which have the knowledge of where each host (VM) is placed. If RBridge edge nodes can utilize the information of where each host is located, then the flooding process to learn the mapping between MAC&VLAN and corresponding RBridge Edge node can be eliminated. This is a great optimization, especially in virtualized data center

environment where VMs migrate all the time. If migrated VMs send out gratuitous ARP (IPv4) or Unsolicited Neighbor Advertisement (IPv6) from the new location, those gratuitous broadcast messages have to flood to all other RBridge edge nodes. If migrated VMs don't send out gratuitous ARP (or ND) from the new location, for packets towards those migrated VMs the ingress RBridge edge nodes will send them to the wrong egress RBridge edge nodes, which is also waste of bandwidth.

The benefits of using directory assistance include:

- The Directory enforced MAC&VLAN <-> RBridge Edge mapping table can determine if a frame needs to be forwarded across RBridge domain.
  - o When multiple Rbridge edge ports are accessible from a server (hosts/VMs), a directory assisted RBridge edge won't flood frames with an unknown DA to all to other RBridge ports. Therefore, there is no need to designate an Appointed Forwarder among all the RBridge Edge ports connected to a Bridge LAN, which enables all RBridge ingress ports to forward traffic.
- Directory assisted approach can not only eliminate the flooding within RBridge domain (unknown learning), but also reduce the flooding on the bridged LAN attached to RBridge edge ports.
- Reduce the amount of MAC&VLAN <-> RBridge edge mapping maintained by RBridge edge. No need for an RBridge edge to keep the MAC entries for hosts which don't communicate with hosts attached to an RBridge edge.

There can be two different models for RBridge edge node to be assisted by Directory:

- Push Model:

Directory Server(s) push down the MAC&VLAN <-> RBridge Edge mapping for all the hosts which might communicate with hosts attached to RBridge edge node.

[Editor's note: there are multiple ways to narrow down the smallest set of remote hosts which communicate with hosts attached to an RBridge edge. A very simple approach: For VLAN #i enabled on one of RBridge Edge port(s), MAC entries for hosts in VLAN #i will not be pushed down to RBridge Edge if

there is no hosts belonging to VLAN #i attached to the RBridge edge. Detailed approaches will be described in a separate draft.]

Whenever there is any change in MAC&VLAN <-> RBridge Edge mapping, which can be triggered by hosts being moved, de-commissioned, or temporarily out of service due to maintenance, an incremental update can be sent to the RBridge edge nodes which are impacted by the change.

Under this model, it is recommended for RBridge edge node to simply drop the data frame (instead of flooding to RBridge domain) if the destination address can't be found in the MAC&VLAN<->RBridge Edge mapping table.

- Pull model:

Under this model, RBridge edge node can simply intercept all ARP requests and forward them to the Directory Server(s) which has the information of how each MAC&VLAN is mapped to its corresponding RBridge edge node.

The reply from the Directory Server can be the standard ARP reply with an extra field showing the RBridge egress node address

RBridge ingress node can cache the mapping

If RBridge edge node receives an unknown MAC-DA, it could choose drop the data frame as in the Push Model, or it can query the directory server. If there is no response from the directory server, the RBridge edge node can drop the frame.

5. Further optimization in using directory assistance for RBridge in data center

The topology between aggregation (or EoR) switches and access (or ToR) switches can be very simple in data center environment. Under those simple topology environments, having the ToR switches participating in RBridge's IS/IS routing domain does not provide much value in topology discovery. By eliminating ToR switches from RBridge routing domain (i.e. let the aggregation (EoR) switches be the boundary of RBridge domain), the number of nodes in the RBridge routing domain can be greatly reduced, which in turns can make the network scale better.

However, two new problems are introduced by letting the aggregation (EoR) switches be the RBridge Domain boundary in the data center environment:

- the number of MAC&VLAN<->RBridge Edge mapping entries to be maintained by the RBridge edge node can be very large (See Scenario #2 in Section 3)
- there is a bridged LAN with multiple ToR switches attached to RBridge Edge port(s), it becomes necessary to have some mechanisms to prevent loops.

The Directory Assistance introduced in this draft (Section 4) provides a solution to the second problem above, i.e. avoid loops among RBridge Edge ports connected to one Bridged LAN so that all RBridge edge ports connected by a Bridged LAN can forward native traffic. However, there is still the problem of too large table of MAC&VLAN <-> RBridge Egress mapping on the edge RBridges.

Therefore, we are proposing further optimization:

- for native Ethernet frames to traverse the RBridge domain, the TRILL encapsulation is done on a node before entering the RBridge domain (e.g. by ToR switches or virtual switch on server), instead of the RBridge Ingress edge node (e.g. EoR switches). That means that the edge ports of the RBridge Ingress node could receive both TRILL-encapsulated data frames and native Ethernet frames. [RBridge] Section 4.6.2 Bullet 8 specifies that an RBridge port can be configured to accept both TRILL encapsulated frames from a neighbor that is not an RBridge.

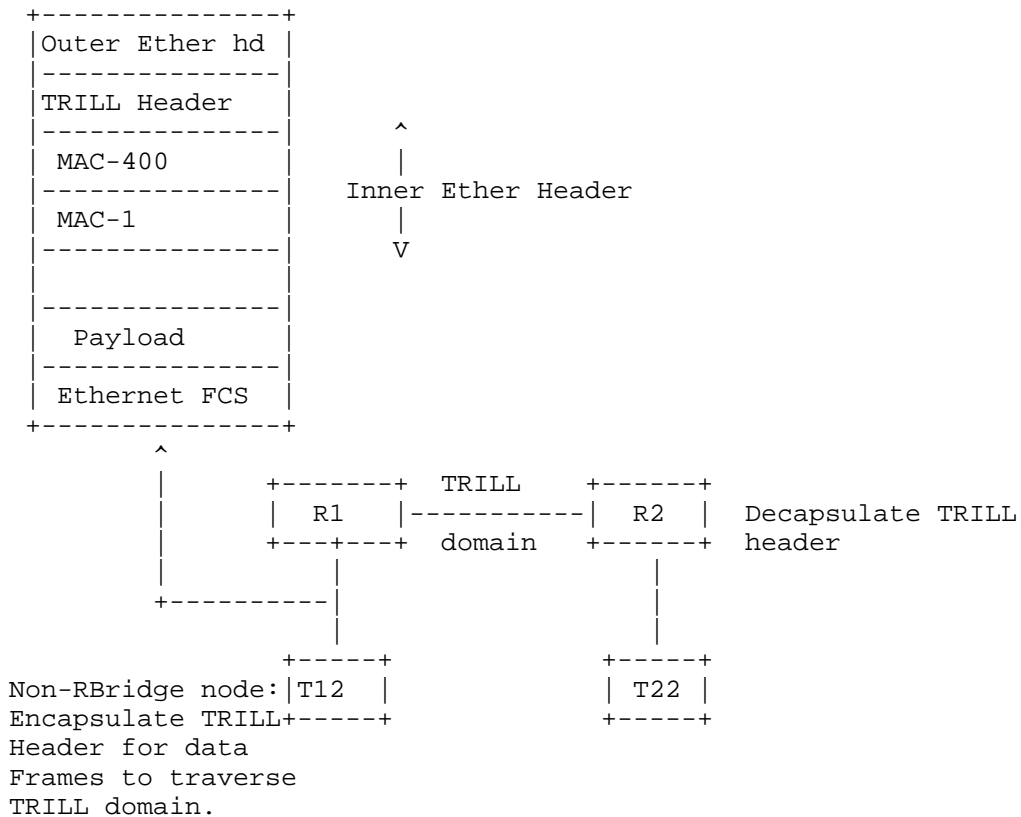
When data frames do not need to traverse RBridge domain, RBridge ingress edge node does its normal native Ethernet data frame processing.

- For egress direction on RBridge edge node, the processing is exactly same as regular RBridge edge node, i.e. decapsulates the TRILL header of the received TRILL frames and forward the decapsulated Ethernet frames to hosts attached to its edge ports.

We call a switch which only performs the necessary TRILL encapsulation for Ethernet data frames to traverse the RBridge domain a "TRILL Encapsulating node" or "Simplified RBridge".



The TRILL Encapsulating Node gets the MAC&VLAN<->RBridge Edge mapping table pushed down or pulled from directory servers. Upon receiving a native Ethernet frame, the TRILL Encapsulating node checks the MAC&VLAN<->RBridge Edge mapping table, and perform the corresponding TRILL encapsulation if the entry is found in the mapping table. If the destination address of the received Ethernet frame and its VLAN doesn't exist in the mapping table, the Ethernet frame is forwarded based on normal Ethernet switching function.



### 5.1. TRILL Header encapsulated by non-RBridge nodes

TRILL header includes Source RBridge's nickname and Destination RBridge's nickname. When a TRILL header is added by a non-RBridge node, using the Ingress RBridge edge node's nickname in the source address field will make the ingress RBridge node receive TRILL

frames with its own nickname in the frames' source address field which can be confusing.

To avoid confusion of Edge RBridges receiving TRILL encapsulated frames with its own nickname in the frames' source address field from neighboring non-RBridge nodes, a new nickname is given to an RBridge edge node, which can be called Phantom Nickname, to represent all the TRILL encapsulating nodes attached to the edge ports of the RBridge edge node.

When the Phantom Nickname is used in the Source Address field of a TRILL frame, it is understood that the TRILL encapsulation is actually done by a non-RBridge node which is attached to an edge port of an RBridge Ingress node.

[Editor's note: a separate draft will be submitted to describe the comprehensive behavior of the ''simplified RBridge'' node]

## 6. Conclusion and Recommendation

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

Therefore, we suggest TRILL to consider directory assisted approach(es). This draft only introduces the basic concept of using directory assisted approach for RBridge edge nodes to learn the MAC&VLAN to RBridge Edge mapping. We want to get some working group consensus before drilling down to detailed steps required for the approach.

## 7. Manageability Considerations

This document does not add additional manageability considerations.

## 8. Security Considerations

TBD.

## 9. IANA Considerations

TBD

## 10. Acknowledgments

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

## 11. References

[RBridges] Perlman, et, al ''RBridge: Base Protocol Specification'',  
<draft-ietf-trill-rbridge-protocol-16.txt>, March, 2010

[RBridges-AF] Perlman, et, al ''RBridges: Appointed Forwarders'',  
<draft-ietf-trill-rbridge-af-02.txt>, April 2011

[ARMD-Problem] Dunbar, et,al, ''Address Resolution for Large Data  
Center Problem Statement'', Oct 2010.

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

## Authors' Addresses

Linda Dunbar  
Huawei Technologies  
1700 Alma Drive, Suite 500  
Plano, TX 75075, USA  
Phone: (972) 543 5849  
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

#### 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](mailto: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: Proposed Standard

Donald Eastlake  
Mingui Zhang  
Huawei  
Puneet Agarwal  
Broadcom  
Dinesh Dutt  
Cisco  
Radia Perlman  
Intel Labs  
July 11, 2011

Expires: January 10, 2012

R Bridges: Fine-Grained Labeling  
<draft-eastlake-trill-rbridge-fine-labeling-01.txt>

## Abstract

The IETF has standardized R Bridges (Routing Bridges), devices that implement the TRILL (Transparent Interconnection of Lots of Links) protocol, a solution for least cost transparent frame routing in multi-hop networks with arbitrary topologies, using link-state routing and encapsulation with a hop count.

The TRILL base protocol standard supports up to 4K VLAN IDs (Virtual Local Area Network Identifiers). However, there are applications that require more fine-grained labeling of data and end stations. This document specifies extensions to the TRILL protocol to 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/lid-abstracts.html>

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

## Table of Contents

1. Introduction.....	3
1.1 Terminology.....	3
2. Fine-Grained Labeling.....	4
2.1 Requirements.....	4
2.2 Existing TRILL VLAN Labeling.....	5
2.3 Fine-Grained Labeling.....	6
3. Coexistence with ST RBridges.....	8
4. Processing Finely Labeled Frames.....	9
4.1 Ingress Processing.....	9
4.2 Transit Processing.....	10
4.2.1 Unicast Transit Processing.....	10
4.2.2 Multi-Destination Transit Processing.....	10
4.3 Egress Processing.....	11
4.4 Address Learning.....	12
5. IS-IS Extensions.....	13
5.1 Announcing RBridge DT Support.....	13
5.2 Interested Labels and Bridge Roots sub-TLV.....	13
5.3 The Group Labeled MAC Address sub-TLV.....	14
6. IANA Considerations.....	16
7. Security Considerations.....	16
7.1 Ingress Forgery and Egress Compromise.....	16
8. References.....	17
8.1 Normative References.....	17
8.2 Informative References.....	17



## 1. Introduction

The IETF has standardized RBridges (Routing Bridges), devices that implement the TRILL (TRansparent Interconnection of Lots of Links) protocol [RFCtrill], a solution for least cost transparent frame routing in multi-hop networks with arbitrary topologies, using link-state routing and encapsulation with a hop count.

The TRILL base protocol standard supports up to 4,094 VLAN IDs (Virtual Local Area Network IDentifiers). However, there are applications that require more fine-grained labeling of data and end stations. This document specifies extensions to the TRILL protocol to accomplish this.

Familiarity with [RFCtrill] and [ISIStrill] is assumed in this document.

### 1.1 Terminology

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

DT - Double Tagging or Double Tagged or Double Tag

Edge RBridge - An RBridge announcing VLAN or fine-grained label connectivity in its LSP

ST - Single Tagging or Single Tagged or Single Tag

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

## 2. Fine-Grained Labeling

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

Section 2.1 lists fine-grained labeling requirements. Section 2.2 briefly outlines VLAN labeling in the TRILL base protocol standard [RFCtrill]. And Section 2.3 then outlines a method of fine-grained labeling of TRILL Data frames.

In the remainder of this document, we commonly refer to the simple VLAN labeling provided by the TRILL base protocol standard as single tagging (ST) or coarse labeling and refer to fine-grained labeling as double tagging (DT).

### 2.1 Requirements

There are several requirements that should be met by fine-grained labeling 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 VLAN tags supported by [RFCtrill] provides for only (  $2^{12} - 2$  ) valid VLAN identifiers. A substantially larger number is required.

#### 2. Silicon Considerations

Fine-grained labeling should, to the extent practical, use existing features, processing, and fields that are already supported in at least some of the many existing TRILL fast path silicon implementations.

#### 3. Base RBridge Compatibility

To support some incremental conversion scenarios, it is desirable that not all RBridges in a campus using fine-grained labeling be required to be fine-grained label aware. That is, it is desirable that RBridges not implementing the fine-grained labeling feature and performing at least the transit forwarding function can

usefully process TRILL Data frames that incorporate fine-grained labeling.

#### 4. Alternate Priority

It would be desirable for an ingress RBridge to be able to assign a different priority to a fine grain labeled 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 Existing TRILL VLAN Labeling

This section provides a brief review of existing TRILL Data frame coarse VLAN labeling.

Currently TRILL Data frames have the single tagged (ST) structure shown below:

```

+-----+
| Link Header |
+-----+
| TRILL Header |
+-----+
| Inner.MacDA  |
+-----+
| Inner.MacSA  |
+-----+
| Inner.VLAN   | <-- Coarse VLAN Label
+-----+
| Payload      |
+-----+
| Link Trailer |
+-----+

```

The Inner.VLAN tag is always present and is specified as a C-tag [802.1Q] providing (  $2^{12} - 2$  ) labels (the values 0 and 0xFFFF are reserved) that is structured as follows:

```

      0 1 2 3 4 5 6 7 8 9 A B C D E F
+---+---+---+---+---+---+---+---+---+---+
|           Ethertype 0x8100           |
+---+---+---+---+---+---+---+---+---+---+
| PRI |C|           VLAN ID           |
+---+---+---+---+---+---+---+---+---+---+

```

The PRI field above is the 3-bit unsigned priority field where larger numbers represent higher priority except that the default zero

priority is above priority 1 and below priority 2 [802.1Q]. Under the TRILL base protocol [RFCtrill], in the Inner.VLAN the C bit is required to be set to zero, transparently forwarded, and ignored on receipt by RBridges.

For an RBridge conformant to the TRILL base protocol, incoming frames are classified as to their VLAN ID and priority by the port on which they are received as described in Appendix D of [RFCtrill].

## 2.3 Fine-Grained Labeling

In the proposed form, fine-grained labeling expands the 12-bit coarse VLAN label available under the TRILL base protocol standard to a 24-bit label. In this document, fine-grained labels are sometimes denoted as "(X.Y)" where X is the high order 12 bits and Y is the low order 12 bits. The fine grained label information appears in the same location in a TRILL Data frame as the coarse VLAN label did, as shown below, although it is encoded as two consecutive VLAN tags (DT).

```

+-----+
| Link Header |
+-----+
| TRILL Header |
+-----+
| Inner.MacDA  |
+-----+
| Inner.MacSA  |
+-----+
| Inner.Label  | <-- Fine-Grained Label
+-----+
| Payload      |
+-----+
| Link Trailer |
+-----+

```

The fine-grained label is encoded as two sequential C-tags as shown below. The high order 12 bits of the fine-grained label appear in the VLAN ID field of the first C-tag and the low order 12 bits appear in the VLAN ID field of the second. Because some silicon might subject the high order part of the fine-grained label to the same constraints as VLAN IDs and for other reasons such as the reporting described in Section 4.2.2, the values zero and 0xFFFF are reserved for the high order part of a TRILL fine-grained label. [[[ Should 0 and 0xFFFF be prohibited in the low order 12 bits also? ]]]

[[[ Alternative Ethertype sequences could be specified. Perhaps the most obvious alternative would be for the first VLAN tag to be as S-tag (Ethertype 0x88A8) and the second a C-tag. However, this might

cause problems for some ST RBridges; if they check the Ethertype of the first VLAN tag, they might reject such frames. ]]]

```

      0 1 2 3 4 5 6 7 8 9 A B C D E F
+-----+-----+-----+-----+
|           Ethertype 0x8100           |
+-----+-----+-----+-----+
| PRI |C| High Order Label Bits |
+-----+-----+-----+-----+
|           Ethertype 0x8100           |
+-----+-----+-----+-----+
| PRI |C| Low Order Label Bits  |
+-----+-----+-----+-----+

```

The appropriate DT for an ingressed native frame is determined by the input RBridge port as specified in Section 4.1. The priority in the second tag is that associated by the ingress port with the native frame as with ST ingress. The priority in the first tag is either a copy of the second tag priority or that priority mapped at ingress, depending on the capabilities of the ingress RBridge. Ports of RBridges supporting DT also have capabilities to transmit frames being forwarded or egressed as untagged or C-tagged as specified in Section 4.3.

Use of S-tags or tags stacked beyond that indicated are beyond the scope of this document but are an obvious extension.

### 3. Coexistence with ST RBridges

ST (single tag) RBridges will operate properly as transit RBridges. Transit RBridges look at the Inner.VLAN ID only for the filtering of multi-destination frames. If an RBridge does not perform filtering, or filters on only some of the fields in the packet, the only consequence is that multi-destination frames will use more bandwidth than necessary. Because ST RBridges could only look at the initial VLAN tag in the fine-grained label of a DT (double tag) multi-destination frame, they will not be able to prune as effectively as transit DT RBridges could.

It would be more serious if an ST edge RBridge, RBl, unaware of the double tag, forwarded a DT frame with DT label (X.Y) onto a link configured as ST VLAN-X, with RBl stripping the "X" and forwarding the packet. This violates the separation of VLANs, and might cause other problems on a link in which the VLAN tag should have been stripped. It would also be problematic if a malicious end station could forge an apparent DT label (X.Y) frame by including extra tags in native frames ingressed by an ST edge RBridge. Therefore, it is highly desirable for all the edge RBridges to be DT RBridges.

DT RBridges will report the DT capability in LSPs, so DT RBridges (and any management system with access to the link state database) will be able to detect the existence of ST edge RBridges.

It might be useful, in a particular campus with mixed DT and ST RBridges, to have some end station VLANs accessible via ST edge RBridges. This is supported by reserving some number of VLANs (say the first k), to be ST-addressable. These VLANs will be specified with a single Inner.VLAN tag, whether or not the edge RBridges attached to these VLANs are DT-capable. When ST-specifiable VLANs are used in a DT campus, and where there are ST edge RBridges advertising connectivity to those VLANs, the first VLAN tag in a double tag MUST NOT be equal to the value of any ST-specifiable VLAN.

If this rule is violated, the network misconfiguration is detected by the DT RBridges that will then refuse in ingress to or egress from label (X.Y) while VLAN X connectivity is being advertised by an ST edge RBridge.

#### 4. Processing Finely Labeled Frames

This section specifies ingress, transit, and egress processing of TRILL Data frames with regard to fine-grained labels, also known as double tagging (DT). A transit or egress DT RBridge detects DT TRILL Data frames by noticing that the Ethertype immediately after the first Inner.Label VLAN tag is the C-tag Ethertype.

##### 4.1 Ingress Processing

There is no change in Appointed Forwarder logic [RFCaf] for the ports of a DT RBridge.

A DT RBridge may be configured, on one or more ports, to double tag ingressed native frames. There is no change in ST ingress processing, which is the default unless a port has been configured for DT.

DT RBridges MUST remove any extra C-tags from incoming native frames being ingressed, regardless of whether the ingress port is configured as ST or DT (see Section 7.1).

DT RBridges MUST support configurable per port mapping from the C-VLAN ID associated with a native frame to a 24-bit fine-grained label. DT RBridges MAY support other methods to determine the DT ID of an incoming native frame. If the resulting label (X.Y) is such that VLAN X connectivity is being advertised by an ST edge RBridge in the campus, the ingressed frame MUST be dropped.

The DT ingress process MUST place the priority associated with an ingressed native frame in the second Inner.Label C-tag. It SHOULD also associate a possibly different mapped priority with an ingressed frame. The mapped priority is placed in the initial Inner.Label C-tag. If such mapping is not supported then the original priority is also placed in the initial inner C-tag.

A DT ingress RBridge MAY serially unicast a multi-destination DT frame to the relevant egress RBridge or RBridges after encapsulating it as a TRILL known unicast data frame. The relevant egress RBridges 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 router connectivity if the native frame was a multicast frame.

## 4.2 Transit Processing

TRILL Data frame transit processing is fairly straightforward as described in Section 4.2.1 for known unicast TRILL Data frames and in Section 4.2.2 for multi-destination TRILL Data frames.

### 4.2.1 Unicast Transit Processing

There is almost no change in TRILL unicast transit processing. A transit RBridge forwards any TRILL unicast data frame to the next hop towards the egress RBridge as specified in the TRILL Header. Just as RBridges conformant to the TRILL base protocol standard [RFCtrill] do not examine the Inner.VLAN ID of ST transit known unicast TRILL Data frames, DT RBridges do not examine either the high or low order part of the 24-bit ID in the Inner.Label for transit DT known unicast TRILL Data frames.

However, as provided in the TRILL base protocol standard [RFCtrill], all transit RBridges, whether ST or DT, MUST take the priority used for a forwarded frame from the Inner.VLAN tag, which will be the first of the two DT VLAN tags for a DT TRILL Data frame.

### 4.2.2 Multi-Destination Transit Processing

All multi-destination TRILL Data frames are forwarded on a distribution tree selected by the ingress RBridge. The distribution trees for DT multi-destination frames are the same trees as for ST multi-destination frames, calculated as provided for in the TRILL base protocol standard [RFCtrill]. There is no change in the Reverse Path Forwarding Check.

A DT RBridge, say RB1, having a DT multi-destination frame for label (X.Y) to forward, SHOULD prune as in the base specification, based on whether there are any edge RBridges on the tree branch that are connected to label (X.Y). In addition, RB1 SHOULD prune multicast frames based on reported multicast listener and multicast router attachment in (X.Y). Finally, a transit DT RBridge MAY drop any multi-destination frame for label (X.Y) if some DT RBridge is advertising connectivity to VLAN X. "MAY" is chosen in this case to minimize the mandatory burden on transit RBridges.

To ensure that a transit ST RBridge does not falsely filter traffic for DT label (X.Y), a DT edge RBridge attached to DT label (X.Y) MUST report connection to VLAN X, as if X were a ST VLAN, in addition to reporting connectivity to label (X.Y). Because of this, DT transit RBridges can safely apply pruning to all TRILL Data frames, both ST



and DT, based on the first Inner.VLAN ID and the reported VLAN-X connectivity of all downstream RBridges.

To ensure that a transit ST RBridge does not falsely prune traffic for DT label (X.Y) base on multicast filtering, a DT edge RBridge attached to label (X.Y) MUST report for VLAN X either (1) that it is attached to both IPv4 and IPv6 multicast routers or (2) its actual DT label (X.Y) multicast listener and router connectivity situation.

#### 4.3 Egress Processing

Egress processing is generally the reverse of ingress processing described in Section 4.1.

If any ST RBridge in the campus is announcing connectivity to VLAN-X, a DT RBridge MUST NOT egress a frame with DT label (X.Y) but must drop such a frame.

A DT RBridge MUST be able to configurably convert the 24-bit fine grained label in a DT TRILL Data frame it is egressing to a 12-bit C-VLAN ID for the resulting native frame on a per port basis. A port MAY be configured to strip such tagging. It is the responsibility of the network manager to properly configure the DT RBridges and ports in the campus to obtain the desired mappings.

A DT RBridge egresses DT frames with the above tag conversion similarly to the egressing of ST frames, as follows:

1. A known unicast DT frame is egressed to the DT port matching its fine-grained label and Inner.MacDA. Or, if there is no such port, it is flooded out all DT ports with its fine-grained label unless the RBridge has knowledge that the frames Inner.MacDA cannot be out that port.
2. A multi-destination DT frame is decapsulated and flooded out all ports with its fine-grained label subject to multicast pruning.

DT RBridges MUST accept multi-destination encapsulated frames that are sent to them as TRILL unicast frames (TRILL Header M bit = 0). They locally egress such frames, if appropriate, and MUST NOT forward them (other than egressing them as native frames on their local links).

#### 4.4 Address Learning

A DT RBridge learns addresses on DT ports based on the fine-grained label rather than VLAN ID. Addresses learned from ingressed native frames are logically represented by { MAC address, fine-grained label, port, confidence, timer } while remote addresses learned from egressing DT frames are logically represented by { MAC address, fine grained label, remote RBridge nickname, confidence, timer }.

## 5. IS-IS Extensions

[[[ Most of the following may be moved to an ISIS draft. ]]]

### 5.1 Announcing RBridge DT Support

An RBridge announces that it is DT in its LSP by ... TBD.

### 5.2 Interested Labels and Bridge Roots sub-TLV

A DT RBridge announces its DT connectivity and related information in the "Interested Labels and Bridge Spanning Tree Roots sub-TLV" (INT-LABEL) which is a variation of the "Interested VLANs and Spanning Tree Roots sub-TLV" (INT-VLAN) structured as below. All fields not defined here are as specified in [ISIStrill].

```

+---+---+---+---+---+
|Type= INT-LABEL|          (1 byte)
+---+---+---+---+---+
|   Length   |          (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Interested Labels   |          (7 bytes)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Appointed Forwarder Status Lost Counter |          (4 bytes)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|           Root Bridges           |          (6*n bytes)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

- o Type: Router Capability sub-TLV Type, set to TBD (INT-LABEL).
- o Length: 13 + 6\*n where n is the number of root bridge IDs.
- o Interested Labels: The Interested Labels field is seven bytes long and formatted as shown below.

```

      0  1  2  3  4  5  6  7
+---+---+---+---+---+---+---+
|M4|M6| R| R| R| R| R| R|
+---+---+---+---+---+---+---+
|                               Label.start - 24 bits                               |
+---+---+---+---+---+---+---+
|                               Label.end - 24 bits                               |
+---+---+---+---+---+---+---+

```

- M4, M6: These bits indicate, respectively, that there is an

IPv4 or IPv6 multicast router on a link for which the originating IS is appointed forwarder for every label in the indicated range.

- R: These reserved bits MUST be sent as zero and are ignored on receipt.
- Label.start and Label.end: This fine-grained label ID range is inclusive. A range of one label ID is indicated by setting them both to that label ID value.

### 5.3 The Group Labeled MAC Address sub-TLV

The existing GMAC-ADDR sub-TLV of the Group Address (GADDR) TLV is specified in [ISIStrill]. It provides for only a 12-bit VLAN-ID. The Group Labeled MAC Address sub-TLV, below, extends this to a 24-bit label.

```

+-----+
|Type=GLMAC-ADDR|                               (1 byte)
+-----+
|   Length      |                               (1 byte)
+-----+
|  RESV  |      Topology-ID      |      (2 bytes)
+-----+
|                24-Bit Label                | (3 bytes)
+-----+
|Num Group Recs |                               (1 byte)
+-----+
|                GROUP RECORDS (1)                |
+-----+
|                .....                |
+-----+
|                GROUP RECORDS (N)                |
+-----+

```

where each group record is of the form:

```

+---+---+---+---+---+
| Num of Sources|                               (1 byte)
+---+---+---+---+---+
|               Group Address           (6 bytes) |
+---+---+---+---+---+
|               Source 1 Address        (6 bytes) |
+---+---+---+---+---+
|               Source 2 Address        (6 bytes) |
+---+---+---+---+---+
|               .....                  |
+---+---+---+---+---+
|               Source M Address        (6 bytes) |
+---+---+---+---+---+

```

- o Type: GADDR sub-TLV Type, set to TBD (GLMAC-ADDR).
- o Length: Variable, minimum 6.
- o RESV: Reserved. 4-bit field that MUST be sent as zero and ignored on receipt.
- o Topology-ID: This field is not currently used in TRILL, where it is sent as zero and ignored on receipt, but is included for use by other technologies.
- o Label: This carries the 24-bit fine-grained label identifier for all subsequent MAC addresses in this sub-TLV, or the value zero if no label is specified.
- o Number of Group Records: A 1-byte integer that is the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. It then has a 48-bit multicast address followed by 48-bit source MAC addresses. If the sources do not fit in a single sub-TLV, the same group address may be repeated with different source addresses in another sub-TLV of another instance of the Group Address TLV.

[[[ Most of the above may be moved to an ISIS draft. ]]]

## 6. IANA Considerations

TBD

## 7. Security Considerations

See [RFCtrill] for general RBridge Security Considerations.

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

More TBD??

### 7.1 Ingress Forgery and Egress Compromise

Confusion between a frame with VLAN-X coarse labeling and DT label (X.Y) is a potential problem.

An end station might try to cause a forged DT TRILL Data frame by sending a double C-tagged frame to a port configured for ST ingress.

The requirement in Section 4.1 that all extra C-tags be removed from native frames on input solves this for DT RBridges. After such removal, the DT RBridge will properly add ST or DT to the encapsulated frame. Thus there is no ingress forgery problem for DT RBridges. However, this does not help for ST RBridges.

ST RBridges need only conform to the [RFCtrill] standard and are not subject to the requirement herein to remove extra C-tags. Thus they might ingress in VLAN-X a native frame double tagged by the end station as (X.Y), removing only the first tag, and then re-insert a VLAN-X tag in the encapsulated frame. The result would be an encapsulated frame that looks like a frame with DT label (X.Y). DT RBridges will think this is a DT frame in (X.Y) and might egress it because they could not distinguish it from a coarsely labeled VLAN-X frame.

Additionally, a TRILL Data frame with DT label (X.Y) could be egressed to VLAN-X by an ST RBridge that is Appointed Forwarder for VLAN-X on one of its ports. Such a frame should not arrive at such an ST RBridge as egress unless the frame is multi-destination.

The above problems are both solved by the prohibition against DT RBridges ingressing to or egressing from DT labeling (X.Y) if the RBridge campus is misconfigured so that an ST edge RBridge is reporting connectivity to VLAN-X while label (X.Y) is in use.

## 8. References

The following sections list normative and informative references for this document.

### 8.1 Normative References

- [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
- [RFCtrill] - R. Perlman, D. Eastlake, D. Dutt, S. Gai, and A. Ghanwani, "RBridges: Base Protocol Specification", draft-ietf-trill-rbridge-protocol-16.txt, in RFC Editor's queue.
- [ISIStrill] - Eastlake, D., A. Banerjee, D. Dutt, R. Perlman, A. Ghanwani, "TRILL Use of IS-IS", draft-ietf-isis-trill-05.txt, in RFC Editor's queue.

### 8.2 Informative References

- [RFCaf] - Perlman, R., D. Eastlake, A. Banerjee, H. Fangwei, "RBridges: Appointed Forwarders", draft-ietf-trill-rbridge-af-03.txt, work in progress.

#### Acknowledgements

The comments and contributions of the following are gratefully acknowledged:

Anoop Ghanwani, Sujay Gupta, Jon Hudson, Vishwas Manral, and Erik Nordmark.

#### 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.3 Xinxu Rd., Shang-Di  
Information Industry Base, Hai-Dian District,  
Beijing, 100085 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

Dinesh G. Dutt  
Cisco Systems  
170 Tasman Drive  
San Jose, CA 95134-1706 USA

Phone: +1-408-527-0955  
Email: ddutt@cisco.com

Radia Perlman  
Intel Labs



INTERNET-DRAFT

RBridges: Fine-Grained Labeling

2200 Mission College Blvd.  
Santa Clara, CA 95054 USA

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

## Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2011 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  
Internet-Draft  
Intended status: Standards Track  
Expires: January 5, 2012

H. Zhai  
F. Hu  
ZTE Corporation  
Radia. Perlman  
Intel Labs  
Donald. Eastlake 3rd  
Huawei technology  
Jul 4, 2011

RBridge: Pseudonode Nickname  
draft-hu-trill-pseudonode-nickname-00.txt

## Abstract

The Appointed Forwarder on a link for VLAN-x is the RBridge that ingresses native frames from the link and egresses native frames to the link in VLAN-x. If the appointed forwarder for an end station is changed, the remote data traffic to the end station could fail. This document is proposed to assign a nickname for pseudonode identifying a multi-access link to solve the issue. When any appointed forwarder encapsulates a packet, it uses the pseudonode nickname as "ingress nickname" rather than its own nickname. If it does, then if the appointed forwarder changes, or the DRB changes, and the pseudonode still uses the same nickname, then the remote RBridge caches won't need to change, and the data traffic to the end station would reach the link uninterruptedly.

## 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 January 5, 2012.

## Copyright Notice

Copyright (c) 2011 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. Problem Statement . . . . .	3
2. Pseudonode Nickname . . . . .	4
3. LSP Announcement . . . . .	4
4. Unicast TRILL Data Frames Processing . . . . .	5
4.1. Ingress processing . . . . .	5
4.2. Egress processing . . . . .	5
4.2.1. Unicasting to VLAN-x Forwarder . . . . .	6
4.2.2. Multicasting to VLAN-x forwarder . . . . .	6
4.2.3. Comparison . . . . .	7
5. TLV Extensions for Pseudonode Nickname . . . . .	8
5.1. Pseudonode Nickname Capability in Hellos . . . . .	8
5.2. Pseudonode Nickname TLV . . . . .	9
5.2.1. Pseudonode Nickname TLV in Hellos . . . . .	10
5.2.2. Pseudonode Nickname TLV in DRB's LSPs . . . . .	10
6. Security Considerations . . . . .	10
7. Acknowledgements . . . . .	10
8. References . . . . .	11
8.1. Normative references . . . . .	11
8.2. Informative References . . . . .	11
Authors' Addresses . . . . .	11

## 1. Problem Statement

The IETF TRILL protocol [RFCtrill] provides optimal pair-wise data frame forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. TRILL accomplishes this by using [IS-IS] [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 R Bridges.

The AF (Appointed Forwarder) on a link for VLAN-x is the R Bridge that ingresses native frames from the link and egresses native frames to the link in VLAN-x. If the appointed forwarder for an end station goes down and a different R Bridge is appointed as appointed forwarder on the link, the end station will not perceive the changes. Therefore, the cache in remote R Bridge could not be correct until it receives the data traffic from the end station, and the traffic from the remote R Bridge to the end station could fail for a while. It is even worse for the Swap Nickname Field approach in multi-level TRILL network, for the egress R Bridge of remote level 1 area cannot update the correspondence of MAC/VLAN-x and the pair of {ingress nickname, swap ingress nickname} until it receives the data traffic from end station [MultilevelTrill].

Pseudonode nickname is proposed in this document to solve the above issue. Pseudonode nickname is assigned by DRB and used to identify a multi-access link. With pseudonode nickname, the data traffic to the end station can reach the destination link uninterruptedly and be forwarded to the end station by other R Bridge even if the appointed forwarder for the VLAN on the link is changed.

The pseudonode nickname is only used in unicast data traffic and not used in multicast data traffic in this document. For the multicast data traffic, the data traffic goes through the distribution tree, and all the R Bridge with the same VLAN can receive the multicast traffic.

This document is organized as following: Section 2 is the concept of pseudonode nickname. Section 3 introduces the LSP announcement mechanism for the pseudonode nickname. Section 4 describes the ingress, transit and egress R Bridge processing of the TRILL data traffic when considering pseudonode nickname. Section 5 specifies pseudonode nickname capability TLV and pseudonode nickname TLV format.

## 2. Pseudonode Nickname

Pseudonode nickname is used to identify a link. It is assigned by DRB on the link. When the RBridge becomes DRB and it doesn't find the pseudonode nickname from TRILL Hello of other RBridges, DRB assigns and announces a pseudonode nickname in its TRILL Hello on the link. If the new DRB obtains the pseudonode nickname from the TRILL Hellos of adjacent RBridges on the link, it reuses this nickname. The nickname for the pseudonode should keep unchanged even if the DRB or AF changed.

All the RBridges on the link should support pseudonode nickname, otherwise the RBridges that don't understand pseudonode nickname on the link cannot forward the encapsulated TRILL frame with pseudonode nickname. Each RBridge on the link announces its pseudonode nickname capability in its TRILL Hello. Only if DRB checks that all the adjacencies in Report state support and enable the pseudonode nickname capability, DRB assigns pseudonode nickname on the link. If not, DRB MUST NOT announce the pseudonode nickname in its pseudonode LSP in the TRILL campus network, otherwise, the remote data traffic may be forwarded to the RBridge without pseudonode nickname capability, and be discarded in the RBridge.

The bypass pseudonode bit is used to determine whether DRB should generate the pseudonode LSP. When bypass pseudonode bit is reset, the DRB should support pseudonode function and generate the pseudonode LSP [TrillAdj]. So if DRB assigns pseudonode nickname on the link, the bypass pseudonode bit MUST be reset in its TRILL Hello.

## 3. LSP Announcement

Pseudonode nickname is only announced in the DRB's pseudonode LSP in the TRILL Network. If one of the RBridges on the link is disabled of the pseudonode nickname function, that is, DRB receives a TRILL Hello without pseudonode nickname capability from the port on the link, the pseudonode nickname function should be disabled on the link, and then DRB updates its pseudonode LSP which doesn't include pseudonode nickname TLV in the TRILL campus network. While if an RBridge (not DRB) supporting pseudonode nickname joins into or exits from the link, it is no influence to the pseudonode nickname LSP originated by DRB. If an RBridge is selected as new DRB and the pseudonode nickname capability on the link is confirmed, it will generate and flood pseudonode LSP including the pseudonode nickname TLV in the TRILL campus network. If DRB finds that the pseudonode nickname function is disabled on the link, it will update its pseudonode LSP which doesn't include pseudonode nickname TLV in the TRILL campus network.

The pseudonode nickname is participated in path computing. The procedure of path computing of pseudonode nickname is same as the routing computing of IPv4 or IPv6 address in layer 3 IS-IS network[RFC1195].

#### 4. Unicast TRILL Data Frames Processing

The processing of TRILL data frames on ingress and egress R Bridges will be influenced when the pseudonode nickname capability is enabled on the link. However, the processing on transit R Bridges remains unchanged.

Section 4.1 covers the changes of processing TRILL data frames on a pseudonode nickname participated ingress R Bridge. Section 4.2 describes two methods to process TRILL data frames on egress R Bridge.

##### 4.1. Ingress processing

When a VLAN-x tagged native frame is sent onto a multi-access link, only the appointed forwarder for that VLAN-x can ingress this frame into TRILL campus. If the pseudonode nickname capability is enabled on the link, the forwarder will encapsulate the frame with a TRILL header, where the ingress nickname is the pseudonode nickname rather than R Bridge's nickname on the link. The encapsulation of the native frame is as same as Section 4.1 in [RFCtrill] except for the ingress nickname in TRILL header.

##### 4.2. Egress processing

On receiving a unicast TRILL data frame, the egress nickname in the TRILL header is examined, and if it is unknown or reserved, the frame is discarded. Then the Inner.VLAN ID, i.e., VLAN-x, is checked. If it is 0x0 or 0xFFFF, the frame is discarded.

This R Bridge will be the egress R Bridge for the TRILL data frame, if the egress nickname is one of the R Bridge's nicknames or one of the pseudonode nicknames of the connected links. If the egress R Bridge is the VLAN-x forwarder on the destination link for this TRILL data frame, the frame is processed and the original self-learning is performed by this R Bridge as described in [RFCtrill]. Otherwise, the frame will be re-encapsulated and transmitted on the link by the egress R Bridge. Only the VLAN-x forwarder can decapsulate the TRILL data frame to native form and forward it to the end station on the link, which is consistent with the principle of ingressing and egressing native frame into and out of TRILL campus, i.e., there is only a single R Bridge on each link that is in charge of ingressing and egressing native frames from and to that link[TrillAdj].



There are two methods for the egress to transmit the re-encapsulated TRILL data frame to VLAN-x forwarder on the link. In section 4.2.1, the egress unicasts the re-encapsulated TRILL data frame to the VLAN-x forwarder, and in 4.2.2, the egress multicasts the TRILL data frame on the link.

#### 4.2.1. Unicasting to VLAN-x Forwarder

To make the final hop, i.e., the egress RBridge (not VLAN-x forwarder), work for a frame addressed to the pseudonode, the forwarding table has to be based on {nickname, VLAN}, instead of {nickname} currently. In the couple of {nickname, VLAN}, nickname is the pseudonode nickname, and VLAN is the VLAN Id of VLAN-x forwarder on this link. If there are several appointed forwarders, each for a VLAN, on this link, several entries exist in the forwarding table, each for a forwarder. In the couple of {nickname, VLAN}, the VLAN will be ignored if the nickname is not a pseudonode nickname on one of local links, and will be set to invalid value (such as 0x0 or 0xFFFF). In other words, if the VLAN in an entry is invalid, the nickname is not a pseudonode nickname.

If the RBridge is not VLAN-x forwarder on the link, it goes to its forwarding table that says, based on the pseudonode nickname and VLAN-x Id, which of its RBridge neighbors, i.e., VLAN-x forwarder on this link, to forward to. The forwarder is identified by the next hop MAC address in the found entry from the above table, which is one of the unicast MAC addresses on one of its ports connected directly on this link. The TRILL data frame is discarded if no entry is found. Otherwise, the outer frame header of the TRILL data frame is stripped, the TRILL header remains unchanged, and a new outer frame header is prepended before the frame is forwarded to the VLAN-x forwarder on the link. For the forwarded frame, the Outer.MacSA is the MAC address of the transmitting port on the destination link, the Outer.MacDA is the next hop MAC address in the found entry and the Outer.VLAN is the designated VLAN on the destination link.

If the above re-encapsulated TRILL data frame is received by a stale VLAN-x forwarder on the destination link, it will be dropped by the RBridge. Otherwise, the re-encapsulated frame is processed as [RFCtrill], and the Inner.MacSA and Inner.VLAN ID are, by default, learned as associated with the ingress nickname unless that nickname is unknown.

#### 4.2.2. Multicasting to VLAN-x forwarder

Alternatively, a special multicast MAC address, named "AF RBridges on this link", can be introduced for the final hop to forward such a TRILL data frame. The scope of the above MAC is limited to local

link, just as the MAC for IS-IS hello PDUs. If a TRILL data frame is addressed to this special MAC and transmitted on a link, all the Appointed Forwarder (AF) RBridges on the link will process it to some extent.

With "AF RBridges on this link", the forwarding table remains unchanged in form, i.e., still based {nickname}. For an entry, the next hop MAC address will be "AF RBridges on this link", if the nickname is the pseudonode nickname on one of local links. In other words, if the nickname is a pseudonode nickname, the next hop MAC MUST be "AF RBridges on this link".

If not VLAN-x forwarder, the final hop RBridge, RBn, looks up its forwarding table, based on the egress nickname in TRILL header of the received frame. The frame will be discarded if no entry is found. Otherwise, RBn will re-encapsulate the frame, i.e., strip the outer frame header, remain the TRILL header unchanged, prepend a new outer frame header before the frame is transmitted onto the link. For the forwarded frame, the Outer.MacSA is one unicast MAC address on the transmitting port connected to the link, the Outer.MacDA is the next hop MAC address in the found entry and the Outer.VLAN is the designated VLAN on the link. If the egress nickname is pseudonode nickname, the Outer.VLAN is "AF RBridges on this link" and the re-encapsulated TRILL data frame is multicasted onto the link.

The TRILL data frame with "AF RBridges on this link" as Outer.MacDA is discarded by other RBridges, which are not AF RBridges, on the link. Otherwise, the Inner.VLAN ID, i.e., VLAN-x, is checked. If the VLAN ID is not valid or the receiving RBridge, RBi, is not VLAN-x forwarder on this link, the frame is also discarded. Else, the TRILL data frame is decapsulated into native form and forwarded to the destination end station, and the Inner.MacSA and Inner.VLAN ID are also, by default, learned as associated with the ingress nickname unless that nickname is unknown by RBi.

#### 4.2.3. Comparison

With the Unicasting method described in Section 4.2.1 above, the re-encapsulated TRILL data frame by the final hop RBridge is only processed by the VLAN-x forwarder on the link, which can reduce the burden of other RBridges as much as possible. But the forwarding table on ingress/egress SHOULD be changed to be based on {nickname, VLAN}, instead of {nickname}, where each AF Rbridge on a local link is identified by the pseudonode nickname and the vlan id of the AF on the link.

With Multicasting method described in Section 4.2.1 above, although all the AF RBridges, except for the final hop RBridge, on the link

are required to process, to some extent, the re-encapsulated TRILL data frame, only the VLAN-x forwarder decapsulates the frame to its native form and forwards it to the destination end station. However, the forwarding table can remain the same as current table in form, i.e., only based on {nickname}.

## 5. TLV Extensions for Pseudonode Nickname

### 5.1. Pseudonode Nickname Capability in Hellos

The Pseudonode nickname capability of an RBridge MUST be included in one subTLV of Port Capability TLV in the RBridge's TRILL Hello PDUs. This capability is included in Special VLANs and Flags (subTLV Type #1) [TrillISIS]. This subTLV MUST appear exactly once in a Port Information TLV in every TRILL Hello PDU. The length of the value is four octets.

Pseudonode Nickname capability TLV

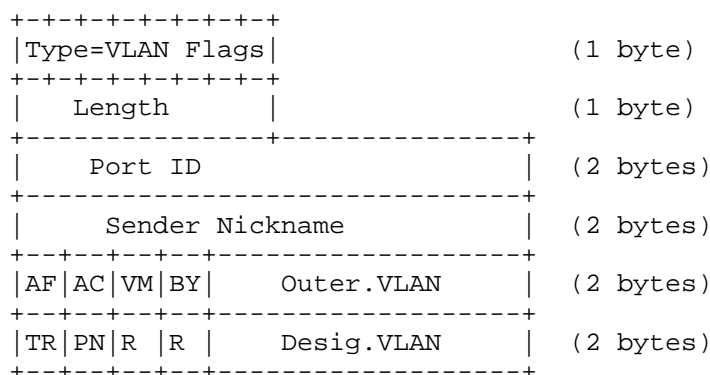


Figure 1

The PN bit, if one, indicates that the sending RBridge supports and enables the pseudonode nickname capability. If an RBridge does not support or not enable this capability, the PN bit MUST be set zero.

Other bits and fields refer to [TrillISIS].

When receiving this subTLV from other RBridges on the link, the DRB can confirm whether all the adjacencies, in Report state [TrillAdj], support and enable this capability. If not, DRB MUST NOT announce pseudonode nickname in its pseudonode LSPs to the TRILL campus, which can avoid the issue that remote traffic is forwarded to a RBridges without pseudonode nickname capability.

## 5.2. Pseudonode Nickname TLV

If the DRB has confirmed that pseudonode nickname capability can be enabled on this link, it will announce the pseudonode nickname to be used on this link in its hello PDUs and in its pseudonode nickname. The pseudonode nickname is carried in Pseudonode Nickname TLV, which is formatted as following:

Pseudonode Nickname TLV

```

+-----+
|Type= PSEU-NICK|                                     (1 byte)
+-----+
|      Length      |                                     (1 byte)
+-----+-----+
|                                     PSEUDONODE NICKNAME RECORDS (1)                                     |
+-----+-----+
|                                     .....                                     |
+-----+-----+
|                                     PSEUDONODE NICKNAME RECORDS (n)                                     |
+-----+-----+

```

where each pseudonode nickname record is of the form:

```

+-----+-----+-----+-----+
|  Nickname.Pri  |SType| Reserved|                                     (2 byte)
+-----+-----+-----+-----+
|                                     Nickname                                     |                                     (2 bytes)
+-----+-----+-----+-----+

```

Figure 2

- o Type: Pseudonode Nickname Type, TBD (NICKNAME).
- o Length: 4\*N, where N is the number of pseudonode nickname records present.
- o SType: An 3-bit unsigned integer sub-type for nickname. If this nickname is pseudonode nickname, value of this field is 1.
- o Nickname.Pri: An 8-bit unsigned integer priority to hold a nickname as specified in Section 3.7.3 of [RFCtrill].
- o Nickname: This is an unsigned 16-bit integer as specified in Section 3.7 of [RFCtrill].

#### 5.2.1. Pseudonode Nickname TLV in Hellos

For an RBridge enabled pseudonode nickname capability on this link, it announces one pseudonode nickname TLV in Hellos if it knows nickname for the pseudonode, otherwise, it MUST NOT announce pseudonode nickname in its Hellos. If DRB has confirmed that pseudonode nickname capability is enabled on this link, the Nickname.Pri in the nickname record MUST be 255, otherwise the Nickname.Pri MUST NOT be 255, and SHOULD be 100 by default.

For an RBridge that is not DRB, it only processes the pseudonode nickname announced by DRB, and MUST overwrite its own pseudonode nickname with the DRB's pseudonode nickname if the two nicknames are different and the Nickname.Pri of DRB is 255. DRB should process the pseudonode nickname TLV from all the adjacencies in the Report state on the link in order to obtain the pseudonode nickname that was being used on this link.

This TLV MUST appear no more than once in a Port Information TLV in every Hello PDU. Only one nickname record can be contained in this TLV, if this subTLV appears in Hello PDUs.

#### 5.2.2. Pseudonode Nickname TLV in DRB's LSPs

For a DRB on a link, it MUST originate and flood a pseudonode LSP for this link if the bypass pseudonode bit is reset. All the adjacencies in the Report state on this link are contained in its pseudonode LSP. Furthermore, if a pseudonode nickname capability is enabled on this link, a Pseudonode Nickname TLV MUST be contained in its pseudonode LSP.

For a pseudonode LSP, the only one record in this TLV contains the nickname for the pseudonode standing for the link. In this case, the value of Nickname.Pri varies from 1 to 255, which describes the DRB's priority to hold this nickname as specified in [RFCtrill] Section 3.7.3.

### 6. Security Considerations

### 7. Acknowledgements

### 8. References

## 8.1. Normative references

## [MultilevelTrill]

Perlman, R., Eastlake, D., and A. Ghanwani, "RBridges: Multilevel TRILL", draft-perlman-trill-rbridge-multilevel-02.txt, work in process, April 2011.

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

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

## [RFCtrill]

Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "RBridges: Base Protocol Specification", draft-ietf-trill-rbridge-protocol-16.txt, in RFC Editor's queue, Mar 2010.

## [TRILLisis]

Eastlake, D., Dutt, D., Perlman, R., and A. Ghanwani, "TRILL Use of IS-IS", draft-ietf-isis-trill-05.txt work in process, Feb 2011.

## [TrillAdj]

Eastlake, D., Perlman, R., Ghanwani, A., Dutt, D., and V. Manral, "RBridges: Adjacency", draft-ietf-trill-adj-02.txt, work in process, Feb 2011.

[TrillAf] Perlman, R., Eastlake, D., Banerjee, A., and F. Hu, "RBridges: Appointed Forwarders", draft-ietf-trill-rbridge-af-03.txt work in process, May 2011.

## 8.2. Informative References

Authors' Addresses

Hongjun Zhai  
ZTE Corporation  
68 Zijinghua Road  
Nanjing 200012  
China

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

Fangwei Hu  
ZTE Corporation  
889 Bibo Road  
Shanghai 201203  
China

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

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 technology  
155 Beaver Street  
Milford, MA 01757  
USA

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





TRILL Working Group  
INTERNET-DRAFT  
Intended status: Proposed Standard  
Updates: RFCtrill

Hongjun Zhai  
Fangwei Hu  
ZTE  
Radia Perlman  
Intel Labs  
Donald Eastlake  
Huawei  
July 3, 2011

Expires: January 2, 2012

RBridges: The ESADI Protocol  
<draft-hu-trill-rbridge-esadi-00.txt>

## Abstract

The IETF TRILL (TRansparent Interconnection of Lots of Links) protocol provides least cost pair-wise data forwarding without configuration in multi-hop networks with arbitrary topologies and safe forwarding even during periods of temporary loops. TRILL supports VLANs and the multi-pathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS (Intermediate System to Intermediate System) link state routing and encapsulating traffic using a header that includes a hop count. Devices that implement TRILL are called RBridges (Routing Bridges).

The ESADI (End System Address Distribution Information) protocol is a VLAN (Virtual Local Area Network) scoped way that RBridge can communicate end station addresses to each other. An RBridge announcing VLAN-x connectivity (normally a VLAN-x forwarder) and running the TRILL ESADI protocol receives remote address information and/or transmits local address information for VLAN-x. The purpose of this document is to improve the documentation of the ESADI protocol. The ESADI RBridge instance states, DRB (Designated RBridge) election procedure, and ESADI sub-TLVs are specified in this document.

## 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: <rbridge@postel.org>.

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

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

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

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

#### Acknowledgements

TBD

## Table of Contents

1. Introduction.....	4
1.1 Content and Precedence.....	5
1.2 Terminology.....	5
2. ESADI Protocol Overview.....	6
3. ESADI DRB Election.....	8
3.1 ESADI DRB.....	8
3.2 ESADI RBridge Instance States.....	8
3.3 ESADI DRB election events.....	9
3.4 Timers.....	9
3.5 ESADI Neighbor List.....	10
3.6 State Table and Diagram.....	10
4. ESADI PDU processing.....	13
4.1 Sending of ESADI PDUs.....	13
4.2 Receipt of PDUs.....	14
5. ESADI LSP Contents.....	16
5.1 ESADI Participation Data.....	16
5.2 ESADI MAC Address sub-TLV.....	17
6. IANA Considerations.....	18
7. Security Considerations.....	18
8. References.....	19
8.1 Normative references.....	19
8.2 Informative References.....	19

## 1. Introduction

The IETF TRILL (TRansparent Interconnection of Lots of Links) protocol [RFCtrill] provides least cost pair-wise data forwarding without configuration in multi-hop networks with arbitrary topologies, 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 (Intermediate System to Intermediate System) [IS-IS] [RFC1195] [TRILLisis] link state routing and encapsulating traffic using a header that includes a hop count. The design supports VLANs (Virtual Local Area Networks) and optimization of the distribution of multi-destination frames based on VLANs and IP derived multicast groups. Devices that implement TRILL are called RBridges (Routing Bridges).

There are five ways an RBridge can learn end station addresses as described in Section 4.8 of [RFCtrill]. The ESADI (End Station Address Distribution Information) protocol is an optional VLAN scoped way RBridges can communicate end station addresses with each other. An RBridge that is announcing connectivity to VLAN-x (normally a VLAN-x appointed forwarder) MAY use the (ESADI) protocol to announce some or all of its attached VLAN-x end nodes.

By default, RBridges with connected end stations learn addresses from the data plane when ingressing and egressing native frames. The ESADI protocols potential advantages over data plane learning include the following:

1. Security advantages: The EDADI protocol can be used to announce end stations with an authenticated enrollment (for example enrollment authenticated by cryptographically based EAP (Extensible Authentication Protocol) methods via [802.1X]). In addition, the ESADI protocol supports cryptographic authentication of its message payloads for more secure transmission.
2. Fast update advantages: ESADI protocol provides a fast update of end nodes MAC (Media Access Control) addresses. If an end station is unplugged from one RBridge and plugged into another one, frames addressed to that older RBridge can be black holed. They can be sent just to the older RBridge that the end station was connected to until cached address information at some remote RBridge times out, possibly for tens of seconds [RFCtrill].

MAC address reachability information and some ESADI parameters are carried in ESADI frames rather than in the core TRILL IS-IS protocol. As described below, ESADI is, for each VLAN, a virtual logical topology overlay in the TRILL topology. An advantage of using ESADI is that the end station attachment information is not flooded to all RBridges through the core IS-IS instance but only to participating RBridges advertising attachment to the VLAN in which those end

stations occur.

### 1.1 Content and Precedence

This document supplements and enriches the description of the ESADI protocol in the TRILL basic specification, especially the ESADI DRB (Designated RBridge) election procedure, ESADI instance state specification, and ESADI parameter announcement. Section 2 is the ESADI protocol overviews. Section 3 specifics ESADI DRB principles, ESADI instance state and DRB election. Section 4 discusses the processing of ESADI PDUs. Section 5 describes two ESADI sub-TLVs: one with ESADI participation information and the MAC Address sub-TLV.

This document updates [RFCtrill] and prevails over [RFCtrill] in the case of conflicts.

### 1.2 Terminology

This document uses the acronyms defined in [RFCtrill].

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

## 2. ESADI Protocol Overview

ESADI is a VLAN scoped way that RBridges can announce and learn end station addresses rapidly and securely. An RBridge that is announcing itself as connected to one or more VLANs (usually because it is an Appointed Forwarder) and participates in the ESADI protocol is called an ESADI RBridge.

ESADI is a separate protocol from the core IS-IS instance implemented by all RBridges in a campus. There is a separate ESADI instance for each VLAN. In essence, for each VLAN, there is an instance of the IS-IS reliable flooding mechanism in which ESADI RBridges may choose to participate. (These are not the instances being specified in [draft-mil].) It is an implementation decision how independent the implementation of multiple ESADI instances at an RBridge are. For example, the ESADI link state could be in a single database with a field in each record indicating the VLAN to which it applies.

After the TRILL header, ESADI frames have an inner Ethernet header with the Inner.MacDA of "All-Egress-RBridges" (formerly called "All-ESADI-RBridges"), an Inner.VLAN tag specifying the VLAN of interest, and the "L2-IS-IS" Ethertype followed by the ESADI payload as shown in Figure 1. For more detail see Section 4.2.5 in the TRILL base protocol specification [RFCtrill].

TRILL ESADI frame Structure

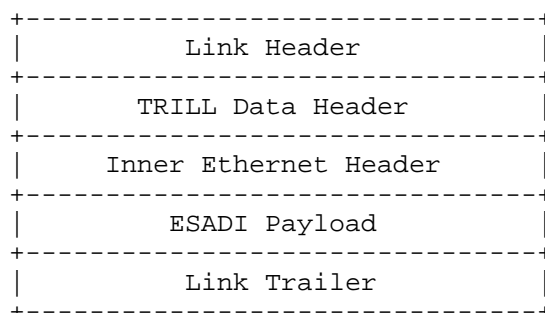


Figure 1

All transit RBridges forward ESADI frames as if they were ordinary multicast TRILL Data frames. Because of this forwarding, it appears to the ESADI protocol at an RBridge that it is directly connected by a multi-access virtual link to all other RBridges in the campus running ESADI for that VLAN. Thus no "routing" computation or decisions ever have to be made by ESADI. A participating RBridge merely transmits the ESADI frames it originates on this virtual link by multicasting it as described in [RFCtrill]. RBridges that do not implement the ESADI protocol, do not have it enabled, or are not

announcing connectivity for the Inner.VLAN of an ESADI frame do not decapsulate or locally process any TRILL ESADI frames they receive. Thus the ESADI frames are transparently tunneled through transit RBridges.

TRILL ESADI frame payloads are structured like IS-IS frames but are always TRILL encapsulated on the wire as if they were TRILL Data frames. The information and adjacency relation between ESADI RBridges are based on the ESADI frames that are carried as TRILL multicast data frames

The ESADI instance for VLAN X at an RBridge acquires a neighbor when it first receives ESADI-LSP fragment zero from that neighbor and that neighbor is an existing RBridge in the core IS-IS instance link state database. When the entry for an RBridge is purged from the core IS-IS link state database, it is also purged from any ESADI instances and is lost as a neighbor.

The information distributed with the ESADI protocol is a list of local end station MAC addresses known to the originating RBridge and, for each such address, a one octet unsigned "confidence" rating in the range 0-254 (see Section 5.2). It is entirely up to the originating RBridge which locally connected MAC addresses it wishes to advertise via ESADI. It MAY advertise all, some, or none of such addresses it has.

TRILL ESADI LSPs do not contain a VLAN ID in their payload. The VLAN ID to which the ESADI data applies is the Inner.VLAN of the frame. If a VLAN ID could occur within the payload, it might conflict with the Inner.VLAN and could conflict with any future VLAN mapping scheme that may be adopted [VLAN-Mapping]. If a VLAN ID field in an ESADI frame payload does include a VLAN ID, its contents is ignored.

### 3. ESADI DRB Election

#### 3.1 ESADI DRB

It is necessary to elect one ESADI RBridge DRB from all the ESADI RBridges for each VLAN where ESADI is being used. The ESADI DRB is responsible for Link State Database synchronization with other RBridges by issuing ESADI-CSNP PDUs periodically and responding to PSNPs on the virtual link.

#### 3.2 ESADI RBridge Instance States

There are four states for each ESADI RBridge instance: Down, Initial, Not-DRB and DRB. The state descriptions are as following:

**Down:** This is a virtual state for convenience in creating state diagrams and tables. It indicates that the ESADI instance is operationally down.

**Initial:** This state indicates that an ESADI instance is up but does not know of any ESADI neighbors (i.e., the only entry in its neighbor list is itself). Once ESADI enters this state, it should start the Holding Timer, and multicast self-originated fragment zero LSPs to other RBridges. If a valid ESADI neighbor is found by receiving an ESADI LSP, the ESADI instance will leave this state and enter into "Not-DRB" state. In this state, the Holding Timer will be reset if the timer is expired or the found neighbor has a higher priority than the local priority.

**Not-DRB:** This state indicates that the ESADI instance has found at least one valid ESADI neighbor and is not DRB yet. If there is no Holding Timer running, the timer will be started. If an ESADI LSP or a CSNP PDU is received from a higher priority ESADI RBridge, the Holding Timer will be reset. If the Holding Timer expires, the ESADI instance will enter into "DRB" state.

**DRB:** In this state, the ESADI instance multicasts the ESADI CSNP PDUs periodically to keep Link State Database synchronization with its neighbors on virtual link, and responds to ESADI-PSNP PDUs with ESADI-LSPs. If an ESADI PDU (i.e., LSP, CSNP and PSNP) is received from a neighbor with a higher priority than its own, the ESADI instance will move to the "Not-DRB" state.



### 3.3 ESADI DRB election events

The following events can change the ESADI state. These are all events for a particular ESADI VLAN-x instance.

E1 ESADI instance is operationally up;

E2 Finding the first ESADI neighbor;

E3 Holding Timer expired;

E4 Receiving an ESADI PDU from an ESADI neighbor with higher priority;

E5 Losing the last ESADI neighbor;

E6 ESADI instance goes operationally down;

(Receiving an ESADI PDU from an ESADI neighbor with lower priority has no effect on the ESADI instance state.)

Priority is determined by the seven-bit priority field in the ESADI participation data (see Section 5.1), with the System ID as a tie breaker, both considered as unsigned integers with the larger quantity indicating higher priority.

### 3.4 Timers

There are two timers for ESADI DRB election: one the Holding Timer, the other the Waiting Timer. The Holding Timer is a cyclic timer, and is used in connection with ESADI-CSNP PDUs. If this timer expires, the local ESADI instance will start multicasting its own ESADI-CSNP PDUs and, if it was in the Non-DRB state, it decides that the DRB is being non-responsive and moves to the DRB state.

The Waiting Timer is a non-cyclic timer. This timer is started by the change of neighbor's DRB status and killed by its expiration. It is used to alleviate the PDUs storm stirred by Link State Database synchronization in the case of current DRB being preempted by a new ESADI neighbor with higher priority. If this timer expires, the new DRB is confirmed and its ESADI parameters, such as intervals of holding timer and waiting timer, are accepted to overwrite the local parameters.

### 3.5 ESADI Neighbor List

In order to be able to access key information about ESADI neighbors easily, an ESADI neighbor list is maintained for each ESADI VLAN-x instance. Each entry in this list represents an ESADI neighbor for VLAN-x.

For each neighbor, there will be a fragment zero LSP from that neighbor in the ESADI instance link state. A list entry is created when such a fragment zero LSP is first received on the ESADI virtual link from some RBridge that exists in the core IS-IS instance link state database. A neighbor entry for an RBridge is deleted when that RBridge is purged from the core IS-IS instance link state database. For each neighbor, the parameters of System-ID/nickname, priority, holding timer interval, waiting timer interval and the DRB flag, are stored in its respective entry in this list.

The DRB flag indicates whether a neighbor is regarded as DRB or not. If this flag is 1, the associated neighbor is considered as DRB, otherwise, not DRB. At any moment, there is no more than one entry that is flagged as DRB in this list. The DRB status of a neighbor can be changed by the receipt of ESADI-CSNP PDUs coupled with the priorities of the originators of the PDUs, together with the priority of the local ESADI instance (see Section 4.2 for more details). When the DRB flag of one entry, such as the entry of the local ESADI instance, is changed in this list, the Waiting Timer will be started if it is not running. When the timer is expired, the neighbor, whose DRB flag is 1, will be confirmed as real DRB, and its CSNP PDUs will be used to accomplish Link State Database synchronization with other ESADI RBridges.

If the ESADI instance is in "Initial" state, there is only one entry existing in this list, where the parameters of the local ESADI VLAN-x instance is saved. If a new entry is added to this list and the entry is the second one, an E2 event will occur, which drives the state of this ESADI instance into "Not-DRB" from "Initial". When there are only two entries in this list, if the second entry is removed from this list, an E5 event is originated, which draws this ESADI instance back to "Initial" state from "Not-DRB" or "DRB".

### 3.6 State Table and Diagram

The table below shows the transitions between the RBridge ESADI instance states defined above based on the events defined above:

TRILL ESADI State Table

Event	Down	Initial	Not-DRB	DRB
E1	Initial	N/A	N/A	N/A
E2	N/A	Not-DRB	N/A	N/A
E3	N/A	Initial	DRB	N/A
E4	N/A	Initial	Not-DRB	Not-DRB
E5	N/A	Initial	Initial	Initial
E6	Down	Down	Down	Down

Figure 2

N/A indicates that the event to the left is Not Applicable in the state at the top of the column.

The first state is "Down". Once an RBridge ESADI instance is operationally up, it enters into "Initial" state. An ESADI instance should start the Holding Timer and multicast its self-originated fragment zero LSP to other ESADI RBridges. When the first valid ESADI neighbor is found on the virtual link, the ESADI instance enters "Not-DRB" state, otherwise the ESADI instance remains in "Initial" state and resets the Holding Timer whenever the timer expires. And if the neighbor's priority is higher than its own, the Holding Timer will be reset before the ESADI instance enters the "Not-DRB" state.

In both the "Non-DRB" and "DRB" states, the ESADI RBridge multicasts all its self-originated LSP fragments.

In the "Not-DRB" state, if any ESADI PDUs are received from ESADI neighbors with higher priorities, the Holding Timer will be reset. Otherwise, if the timer expires without hearing from a higher priority neighbor, the ESADI instance will enter "DRB" state. An ESADI DRB can be preempted by a higher priority neighbor. If the DRB receives an ESADI PDU from a higher priority neighbor, the ESADI instance will move to "Not-DRB" state. As DRB, an ESADI instance will multicast ESADI-CSNP PDUs to all neighbors on the virtual link periodically, and respond to the ESADI-PSNP PDUs with ESADI-LSP PDUs by multicasting them.

Below is the same information as in the state table presented as a diagram.

TRILL ESADI state diagram

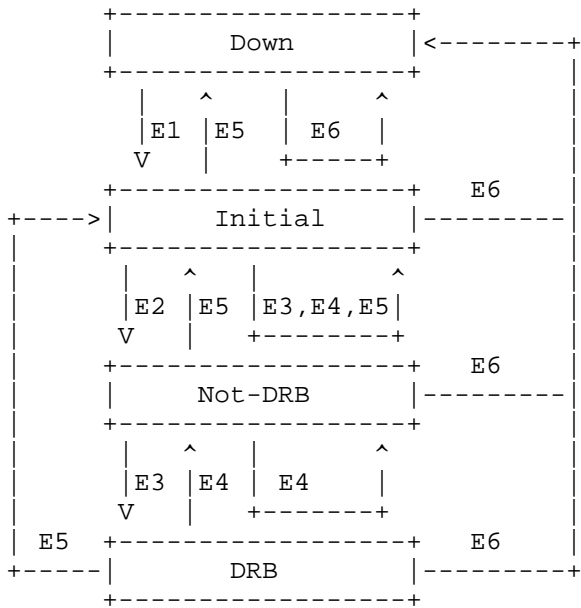


Figure 3

#### 4. ESADI PDU processing

VLAN-x ESADI neighbors are usually not connected directly by a physical link, but are always logically connected by a virtual link. There may be hundreds of ESADI RBridges on the virtual link. There are only LSP, CSNP and PSNP PDUs used in ESADI. In particular, there are no Hellos because ESADI does not need to build a topology and does not need to do any routing.

In IS-IS, multicasting is normally on a local link and no effort is made to optimize to unicast because under the original conditions when IS-IS was designed (commonly a piece of multi-access Ethernet cable), any frame made the entire link busy for that frame time. But in ESADI what appears to be a simple multi-access link is actually a multi-hop distribution tree that may or may not be pruned. Thus, transmitting a multicast frame on such a tree imposes a substantially greater load than transmitting a unicast frame. This load may be justified if there are likely to be multiple listeners but may not be justified if there is only one recipient of interest. For this reason, under some circumstances, ESADI PDUs MAY be unicast.

An undesirable storm of LSP PDUs may be sent to update a new ESADI RBridge when it starts to participate in VLAN-x ESADI if it has higher priority and becomes DRB on the virtual link.

Section 4.1 describes the sending of ESADI PDUs. Section 4.2 covers the receipt of ESADI PDUs.

##### 4.1 Sending of ESADI PDUs

When the VLAN-x ESADI instance is in "Not-DRB" or "DRB" state and a new neighbor is found, its self-originated LSP fragments are scheduled to be sent and MAY be unicast to that neighbor. The interval elapsed before sending the LSP(s), depends on the priority of the local ESADI instance. The higher the priority, the shorter the interval is.

In the case of receiving an LSP with a smaller sequence number than the LSP copy stored in local Link State Database, the local ESADI instance will also schedule to transmit the stored LSP copy and MAY unicast it to the sender. After the sender receives such a LSP, it can originate a new LSP, whose sequence number is bigger than the received sequence number, to refresh the LSP in all the neighbors.

If the ESADI instance is DRB, it multicasts a CSNP periodically to keep the Link State Database synchronized among its neighbors on the virtual link. After receiving a PSNP PDU, the DRB will transmit the

LSPs requested by the PSNP on the virtual link.

If the ESADI instance is not DRB, it will schedule multicasting only its self-originated LSP on the virtual link when it finds the DRB losing some LSPs or having stale LSPs, including the local ESADI instance's self-originated LSPs, from the CSNP PDUs it receives. The higher the priority, the shorter the interval is.

The format of a unicast ESADI frame is the format of TRILL ESADI frame, in section 4.2 in [RFCtrill], except that, in the TRILL header, the M bit is set to zero and the Egress Nickname is the nickname of the destination RBridge.

#### 4.2 Receipt of PDUs

When an ESADI PDU is received, the receiver checks for the originator's System ID in the receiver's core IS-IS instance link state database. If the System ID is not present, the PDU is discarded.

After receiving a new ESADI LSP with a known System ID, the LSP will be installed into or replaced the older copy of this LSP in the local ESADI Link State Database. If it is a fragment zero LSP, the local ESADI instance will try to find the originator of the LSP in its neighbor list. If the neighbor is found, any different parameters of this neighbor will be stored in the associated entry in the neighbor list. Otherwise, a new neighbor is detected, and an associated entry is inserted into the list to store this neighbor's information. If the local ESADI instance is in "Initial" state and the entry is the second entry in the neighbor list, an E2 event will be originated, which will push this ESADI instance into "Not-DRB" state from "Initial".

On receiving a CSNP PDU with a known System ID, if the ESADI instance is in "Initial" state, the PDU is used for Link State Database synchronization. Otherwise, the local ESADI instance tries to find the originator in its neighbor list. In the following cases, the CSNP PDU will be discarded:

- 1) The originator is not found in the neighbor list;
- 2) The originator is found in the list, but its priority is not higher than the local ESADI instance's priority.

If the CSNP PDU is not discarded and the local ESADI instance is in "Not-DRB" or "DRB" state, it will be used to update the DRB flag in the entries in the neighbor list, i.e., the originator's DRB flag is set to 1 and other entries' DRB flag is cleared to zero. The above update may make the potential DRB change from one neighbor to

another; if so, the Waiting Timer will be started if it is not running. If potential DRB doesn't change from one neighbor to another, and the Waiting Timer is not running yet, this CSNP PDU is used for Link State Database synchronization. ESADI-PSNP PDUs will be multicast on the virtual link to request fresh copies of lost or stale LSPs from DRB, if necessary.

When receiving a PSNP PDU, if the local ESADI instance is DRB and the Waiting Timer is not running, LSP PDU associated the PSNP will be multicast on the virtual link. Otherwise, the PSNP PDU is discarded.

## 5. ESADI LSP Contents

The only PDUs used in ESADI are the Level 1 LSP, CSNP, and PSNP PDUs. This section specifies the format for ESADI participation data sub-TLV and gives the reference for the ESADI MAC Reachability TLV.

### 5.1 ESADI Participation Data

The figure below presents the format of the ESADI participation data. This sub-TLV MUST be included in LSP fragment zero. LSP fragment zero MUST NOT exceed 1470 bytes in length.

[[ This should probably be a Router Capabilities TLV and this section probably needs to be in an ISIS WG draft. ]]

#### Participation Data

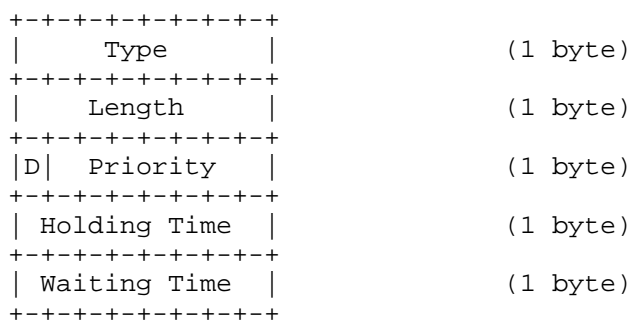


Figure 4

Type: set to TBD

Length: Set to 3.

D: If the sub-TLV is originated by the DRB or an EASDI instance sending CSNPs as DRB, the D field is set to 1, otherwise, the field is zero.

Priority: The Priority field gives the RBridge's priority for being DRB on the TRILL EASDI virtual link for the VLAN in which the PDU containing the Participation data was sent. It is a 7-bit unsigned integer.

Holding Time: Gives the holding time in seconds as an unsigned integer.



Waiting Time: Gives the waiting time in seconds as an unsigned integer.

When an ESADI instance receives a participation data sub-TLV in which the D field is set to 1 and the originator of this LSP is confirmed DRB by the local ESADI instance as highest priority, the RBridge sets the local Holding Timer according to the value of holding time field, and sets the local Waiting Timer according to the value of waiting time field.

## 5.2 ESADI MAC Address sub-TLV

The information in TRILL ESADI LSP PDUs consists of one or more MAC Reachability (MAC-RI) TLVs as specified in [RFC6165]. These TLVs contain one or more unicast MAC addresses of end stations that are both on a port and in a VLAN for which the originating RBridge is appointed forwarder, along with the one octet unsigned Confidence in this information with a value in the range 0-254.

To avoid conflict with the Inner.VLAN ID, the TLVs in TRILL ESADI PDUs, including the MAC-RI TLV, MUST NOT containing the VLAN ID. If a VLAN-ID is present in the MAC-RI TLV, it is ignored. The VLAN to which the ESADI-LSP applies is indicated only by the Inner.VLAN tag in the encapsulated TRILL ESADI frame.

## 6. IANA Considerations

TBD

## 7. Security Considerations

TBD

For general TRILL Security Considerations, see [RFCtrill].

## 8. References

Normative and informative references for this document are below.

### 8.1 Normative references

- [IS-IS] - International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, Nov 2002.
- [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.
- [RFC4971] - Vasseur, JP., N. Shen, and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [RFC6165] - Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.
- [TRILLisis] - Eastlake, D., A. Banerjee, D. Dutt, R. Perlman, A. Ghanwani, "TRILL Use of IS-IS", draft-ietf-isis-trill-05, in RFC Editor's queue.
- [RFCtrill] - Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "RBridges: Base Protocol Specification", draft-ietf-trill-rbridge-protocol-16.txt, in RFC Editor's queue, Mar 2010.

### 8.2 Informative References

- [802.1X] - IEEE 802.1, "IEEE Standard for Local and metropolitan area networks / Port-Based Network Access Control", IEEE Std 802.1X-2004, 13 December 2004.
- [VLAN-Mapping] - Perlman, R., D. Dutt, A. Banerjee, A. Rijhsinghani, and D. Eastlake, "RBridges: Campus VLAN and Priority Regions", draft-ietf-trill-rbridge-vlan-mapping-05.txt, work in process, April 2011.

Authors' Addresses

Hongjun Zhai  
ZTE Corporation  
68 Zijinghua Road  
Nanjing 200012 China

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

Fangwei Hu  
ZTE Corporation  
889 Bibo Road  
Shanghai 201203 China

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

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  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757 USA

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

## Copyright and IPR Provisions

Copyright (c) 2011 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: Proposed Standard

Donald Eastlake  
Huawei  
Anoop Ghanwani  
Brocade  
Vishwas Manral  
IP Infusion  
Caitlin Bestler  
Quantum  
July 10, 2011

Expires: January 9, 2012

R Bridges: TRILL Header Extensions  
<draft-ietf-trill-rbridge-options-05.txt>

## Abstract

The TRILL base protocol standard specifies minimal hooks to safely support TRILL Header extensions. This draft specifies the format for such extensions and specifies some initial extensions.

## 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 <rbridge@postel.org>.

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

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

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

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

## Table of Contents

1. Introduction.....	3
1.1 Conventions used in this document.....	3
2. TRILL Header Options.....	4
2.1 RBridge Extension Handling Requirements.....	5
2.2 No Critical Surprises.....	6
2.3 Extensions Format.....	6
2.3.1 Extended Header Flags Area.....	7
2.3.1.1 Critical Summary Bits.....	8
2.3.1.2 MEF, More Extended Flags.....	8
2.3.1.3 Specific Initial Bit Extended Flags.....	9
2.3.1.4 TLV Summary Bits.....	9
2.3.1.5 Flow ID.....	9
2.3.2 TLV Extension Format.....	10
2.3.3 Marshaling of Extensions.....	11
2.4 Conflict of Extensions.....	11
3. Specific Extended Header Flag.....	13
3.1 The Alert Extended Flag.....	13
3.2 The ECN Extension.....	13
4. Specific TLV Extension.....	16
4.1 Test/Pad Extension.....	16
5. Additions to IS-IS.....	17
6. IANA Considerations.....	18
7. Security Considerations.....	18
8. Acknowledgements.....	18
9. References.....	19
9.1 Normative References.....	19
9.2 Informative References.....	19
Change History.....	20



## 1. Introduction

The base TRILL protocol standard [RFCtrill] provides a TRILL Header extensions feature, called "options" in [RFCtrill], and describes minimal hooks to safely support header extensions. But, except for the first two bits, it does not specify the structure of the extension to the TRILL Header nor the details of any particular extension. This draft specifies that format and some initial extensions: a special Flow ID field, ECN (Explicit Congestion Notification) extended header flags, an Alert extended header flag, and a test/pad extension.

Section 2 below describes the general principles, format, and ordering of TRILL Header Extensions. Other than the special Flow ID extension, TRILL Header extensions are of two kinds: extended header flags and TLV (Type, Length, Value) encoded extensions.

Section 3 describes two specific extended flag extensions while Section 4 describes a specific TLV encoded extension.

### 1.1 Conventions used in this document

The terminology and acronyms defined in [RFCtrill] are used herein with the same meaning.

In this documents, "IP" refers to both IPv4 and IPv6.

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

## 2. TRILL Header Options

The base TRILL Protocol includes a feature for extension of the TRILL Header (see [RFCtrill] Sections 3.5 and 3.8). The 5-bit Op-Length header field gives the length of the extension to the TRILL Header in units of 4 octets, which allows up to 124 octets of header extension. If Op-Length is zero there is no header extension present; else, this area follows immediately after the Ingress Rbridge Nickname field of the TRILL Header. The optional extensions area consists of an extended flags area possibly followed by TLV extensions. Each TLV extension present is 32-bit aligned. There is a special Flow ID extension that may also occur in the extended flags area.

As described below, provision is made for both hop-by-hop extensions, which might affect any RBridge that receives a TRILL Data frame containing such an extension, and ingress-to-egress extensions, which would only necessarily affect the RBridge(s) where a TRILL frame is decapsulated. Provision is also made for both "critical" and "non-critical" extensions. Any RBridge receiving a frame with a critical hop-by-hop extension that it does not implement MUST discard the frame because it is unsafe to process the frame without understanding the critical extension. Any egress RBridge receiving a frame with a critical ingress-to-egress extension it does not implement MUST drop the frame if it is a known unicast frame; if it is a multi-destination TRILL Data frame with a critical ingress-to-egress extension that the RBridge does not implement, then it MUST NOT be egressed at that RBridge but it is still forwarded on the distribution tree. Non-critical extensions can be safely ignored.

Any extension indicating a significant change in the structure or interpretation of later parts of the frame which, if the extension were ignored, could reasonably cause a failure of service or violation of security policy MUST be a critical extension. If such an extension affects any fields that transit RBridges will examine, it MUST be a hop-by-hop critical extension.

TLV extensions also have a "mutability" flag that has a different meaning for ingress-to-egress and for hop-by-hop.

For an ingress-to-egress extension, the mutability flag indicates whether the value associated with the extension can change at a transit RBridge (mutable extensions) or cannot so change (immutable extensions). For example, an ingress-to-egress security extension could protect the value of an immutable ingress-to-egress extension. But such a security extension generally could not protect a mutable value as a transit RBridge could change that value but might not have the keys to recompute a signature or authentication code to take a changed value into account.

For a non-critical hop-by-hop extension, the mutability flag

indicates whether a transit RBridge that does not implement the extension is permitted (mutable) or not permitted (immutable) to remove the extension. A transit RBridge is not required to remove a hop-by-hop extension that it does not implement.

For critical hop-by-hop extensions, the mutability flag is meaningless. If the RBridge does not implement the critical hop-by-hop extension, it MUST drop the frame. If it does implement the critical hop-by-hop extension, it will know whether or not it may/should/must remove it. For critical hop-by-hop extensions, the mutability flag is set to zero ("immutable") on transmission and ignored on receipt.

Note: Most RBridge implementations are expected to be optimized for simple and common cases of frame forwarding and processing. Although the hard limit on the header extensions area length, the 32-bit alignment of TLV extensions, and the presence of critical extension summary bits, as described below, are intended to assist in the efficient hardware based processing of frames with a TRILL header extensions area, nevertheless the inclusion of extensions, particularly TLV extensions, may cause frame processing using a "slow path" with inferior performance to "fast path" processing. Limited slow path throughput of such frames could cause them to be discarded.

## 2.1 RBridge Extension Handling Requirements

The requirements given in this section are in addition to the extension handling requirements in [RFCtrill] (where they are called options).

All R Bridges MUST be able to check whether there are any critical extensions present that are necessarily applicable to their processing of the frame as detailed below. If they do not implement all such critical extensions present, they MUST discard the frame or, in some circumstances as described above for certain multi-destination frames, continue to forward the frame but MUST NOT egress the frame.

Transit R Bridges MUST transparently forward all immutable ingress-to-egress header extensions in frames that they forward. Any changes made by a transit RBridge to a mutable ingress-to-egress extension value MUST be a change permitted by the specification of that extension.

In addition, a transit RBridge:

- o MAY add, if space is available, or remove hop-by-hop extensions as

- specified for such extensions;
- o MAY change the value and/or length of a mutable ingress-to-egress TLV extension as permitted by that extension's specification and provided there is enough room if lengthening it;
  - o MUST adjust the length of the extensions area, including changing Op-Length in the TRILL header, as appropriate for any changes it has made;
  - o MUST NOT add, remove, or re-order ingress-to-egress extensions.
  - o with regard to any non-critical hop-by-hop extensions that the transit RBridge does not implement, it MAY remove them if they are mutable but MUST transparently copy them when forwarding a frame if they are immutable.

## 2.2 No Critical Surprises

RBridges advertise the ingress-to-egress extensions they support in their IS-IS LSP and advertise the hop-by-hop extensions they support at a port on the link connected to that port. An RBridge is not required to support any extensions.

Unless an RBridge advertises support for a critical extension, it will not normally receive frames with that extension.

An RBridge SHOULD NOT add a critical extension to a frame unless,

- for a critical hop-by-hop extension, it has determined that the next hop RBridge or RBridges that will accept the frame support that extension, or
- for a critical ingress-to-egress extension, it has determined that the RBridge or RBridges that will egress the frame support that extension.

"SHOULD NOT" is specified since there may be cases where it is acceptable for those frames, particularly for the multi-destination case, to be discarded by any RBridges that do not implement the extension.

## 2.3 Extensions Format

If any extensions are present in a TRILL Header, as indicated by a non-zero Op-Length field, the first 32 or 64 bits of the extensions area consist of extended header flags and the Flow ID, as described below. The remainder of the extensions area, if any, after this initial 32 or 64 bits, consists of TLV (Type Length Value) extensions aligned on 32-bit boundaries. Section 2.3.2 specifies the format of a TLV extension. Section 2.3.3 describes the marshaling of TLV extensions.

### 2.3.1 Extended Header Flags Area

The first 32 bits of the Extensions Area are organized as follows:

	0	1	2	3-4	5-7	8-10	11-12	13	14	15		16 - 31										
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----									
	CHbH		CItE		MEF		CHHF		NHHF		CIEF		NIEF		NHHT		CIET		NIET		Flow ID	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

Figure 1: Extensions Area Initial 32 Bits

Any RBridge adding an extensions area to a TRILL Header must set these 32 bits to zero except when permitted or required to set one or more of them as specified. The meanings of these bits are listed in the table below and then further described.

Bit(s)	Description
0	CHbH: Critical Hop-by-Hop extension(s) are present.
1	CItE: Critical Ingress-to-Egress extension(s) are present.
2	MEF: More Extended Flags, indicates that an additional 32-bit extended flags area is present as described below.
3-4	CHHF: Critical Hop-by-Hop extended Flag bits.
5-7	NHHF: Non-critical Hop-by-Hop extended Flag bits.
8-10	CIEF: Critical Ingress-to-Egress extended Flag bits.
11-12	NIEF: Non-critical Ingress-to-Egress extended Flag bits.
13	NHHT: Non-critical Hop-by-Hop TLV extension(s) are present.
14	CIET: Critical Ingress-to-Egress TLV extension(s) are present.
15	NIET: Non-critical Ingress-to-Egress TLV extension(s) are present.
16-31	Flow ID if non-zero.

All extended flags are considered mutable except the critical hop-by-hop extended flags.

For TRILL Data frames with extensions present, any transit RBridge MUST transparently copy bits 8 through 12, except as permitted by an extension implemented by that RBridge, but MAY either copy or clear any of the bits 5 through 7. Even if a transit RBridge removes all TLV extensions from a TRILL Header when allowed to do so, it MUST NOT eliminate the extensions area in a forwarded frame if any of bits 3, 4, or 8 through 12 remain non-zero; however, if there are no TLV extensions and all of bits 2 through 31 are zero, then the summary bits will also be zero and the transit RBridge MAY eliminate the Extensions area in the frame, setting Op-Length to zero.

### 2.3.1.1 Critical Summary Bits

The top two bits of the extensions area, bits 0 and 1 above, are called the critical summary bits. They summarize the presence of critical extensions as follows:

**CHbH:** If the CHbH (Critical Hop by Hop) bit is one, one or more critical hop-by-hop extensions are present in the extensions area. Transit RBridges that do not support all of the critical hop-by-hop extensions present, for example an RBridge that supported no hop-by-hop extensions, **MUST** drop the frame. If the CHbH bit is zero, the frame is safe, from the point of view of extensions processing, for a transit RBridge to forward, regardless of what extensions that RBridge does or does not support. A transit RBridge that supports none of the extensions present **MUST** transparently forward the extensions area when it forwards a frame, except that it **MAY** remove mutable hop-by-hop extensions.

**CItE:** If the CItE (Critical Ingress to Egress) bit is a one, one or more critical ingress-to-egress extensions are present in the extensions area. If it is zero, no such extensions are present. If either CHbH or CItE is non-zero, egress RBridges that do not support all critical extensions present, for example an RBridge that supports no extensions, **MUST** drop the frame. If both CHbH and CItE are zero, the frame is safe, from the point of view of extensions, for any egress RBridge to process, regardless of what extensions that RBridge does or does not support.

The critical summary bits enable efficient processing of TRILL Data frames by RBridges that support no critical extensions and by transit RBridges that support no critical hop-by-hop extensions. Such RBridges need only check whether Op-Length is non-zero and, if it is, the top one or two bits just after the fixed portion of the TRILL Header.

### 2.3.1.2 MEF, More Extended Flags

Bit 2, if set, indicates there are an additional 32 bits of extended flags. They are organized as shown below. The start of the TLV extensions, if any, is moved to after these additional bit extensions.

	32 - 39		40 - 47		48 - 55		56 - 63	
+	+	+	+	+	+	+	+	+
	Critical HbH		NonCritical HbH		Critical ItE		NonCritical ItE	
+	+	+	+	+	+	+	+	+

Figure 2: Extended Flag Bits 32 to 63

### 2.3.1.3 Specific Initial Bit Extended Flags

CHHB, bits 3 and 4, are Critical Hop-by-Hop Bits.

NHHB, bits 5 through 7, are Non-critical Hop-by-Hop Bits.

CIEB, bits 8 through 10, are Critical Ingress-to-Egress Bits.

NIEB, bits 11 and 12, are Non-critical Ingress-to-Egress Bits.

The bits above are available for indicating extended header flags, except for the bits allocated by Section 3 below.

### 2.3.1.4 TLV Summary Bits

It is anticipated that in most cases the interpretation of TLV encoded extensions in TRILL data frames will be handled by slow path software. To minimize unnecessary resort to the slow path, the TLV summary bits, plus a special check for critical hop-by-hop TLV extensions, enable an RBridge to quickly determine if any TLV encoded extensions of the category or categories it implements are present.

Bits 13-15, the NHHT, CIET, and NIET bits, indicate the presence later in the TRILL Header of TLV encoded Non-critical Hop-by-Hop, Critical Ingress-to-Egress, and Non-critical Ingress-to-Egress TLV extensions respectively.

There is no Critical Hop-by-Hop TLV flag bit because the presence of one or more such TLV extensions can be determined by examining Op-Length and, if Op-Length and the MEF bit indicate that there are TLV extensions beyond the extended flags area, examining the top two bits of the first extensions area byte after the extended flags area. The ordering restrictions on TLV extensions require that, if any Critical Hop-by-Hop TLV extensions are present, they appear first in the TLV extensions area. Thus it is adequate to check only if the first TLV extension present is a Critical Hop-by-Hop extension, which can be determined from the top two bits of its first byte.

### 2.3.1.5 Flow ID

In connection with the multi-pathing of frames, frames that are part of the same order-dependent flow need to follow the same path. Methods to determine flows are beyond the scope of this document; however, it may be useful, once the flow of a unicast frame has been determined, to preserve and transmit that information for use by subsequent RBridges.

The Flow ID extension is a specially encoded non-critical hop-by-hop extension that appears in bits 16 through 31 of the initial bit encoded extensions area. Its presence is indicated by a non-zero value in that field.

It is considered hop-by-hop because it can be added or changed by a transit RBridge and transit RBridges can use it to make forwarding decisions. Because the ingress RBridge may know the most about a frame, it is expected that this extension would most commonly be added at the ingress RBridge. Once set non-zero in a frame, the extension SHOULD NOT be removed, set to zero, or changed unless, for example, a campus is divided into regions such that different Flow IDs would make sense in different regions.

### 2.3.2 TLV Extension Format

TRILL Header extensions, other than the extended header flags and Flow ID described above, are TLV encoded, with some flag bits in the Type and Length octets, in the format show in Figure 3.

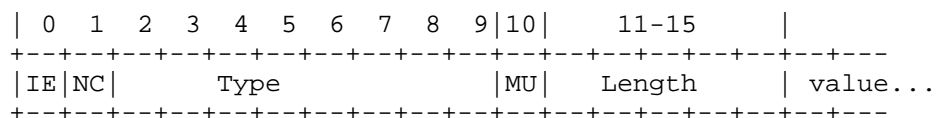


Figure 3. Extension TLV Structure

The highest order bit of the first octet (IE) is zero for hop-by-hop extensions and one for ingress-to-egress extensions. Hop-by-hop extensions are potentially applicable to every RBridge that receives the frame. Ingress-to-egress extensions are only inserted at the ingress RBridge and are applicable at egress RBridges. Ingress-to-egress extensions MAY also be examined and acted upon by transit RBridges as specified in the particular extension.

The second highest order bit of the first octet (NC) is zero for critical extensions and one for non-critical extensions.

Bit 10 in the second octet (MU) is zero for immutable extensions and one for mutable extensions. The IE, NC, Type, and MU fields themselves MUST NOT be changed even for a mutable extension.

The eight-bit Type code extends from bit 2 through bit 9. The extension Type may constrain the values of the IE, NC, and MU bits. For example, a certain Type might require that the extension be marked as a hop-by-hop, non-critical, mutable extension. If the IE, NC, or MU bits have a value not permitted by the extension Type specification for an extension that an RBridge must act on (any



critical ingress-to-egress extension at an egress RBridge and any critical hop-by-hop extension), the RBridge MUST discard the frame. If these bits have a value not permitted by for the Type for an extension that an RBridge may ignore (any ingress-to-egress extension at a transit RBridge and any non-critical extension), the RBridge MAY discard the frame. "MAY" is chosen in this case to minimize the checking burden.

The Length field is an unsigned quantity giving the length of the extension value in units of four octets. It gives the size of the extension including the initial two Type and Length octets. The Length field MUST NOT be such that the extension value extends beyond the end of the total extensions area as specified by the TRILL Header Op-Length. Thus, the value 31 is reserved and, when such a value is noticed in a frame, the frame MUST be discarded.

### 2.3.3 Marshaling of Extensions

In a TRILL Header with extensions, those extensions start immediately after the Ingress RBridge Nickname and fill the extensions area. TLV extensions are 32-bit aligned.

TLV extensions start immediately after the initial four or eight octets of extended flags area and MUST appear in ascending order by the value of the eleven high order bits (bits 0 through 10) of the Type and Length octets considered as an unsigned integer in network byte order. There MUST NOT be more than one extension in a frame with any particular value of this eleven high order bits. Thus the TLV extensions MUST be ordered as follows: (1) critical hop-by-hop extensions, (2) non-critical hop-by-hop extensions, (3) critical ingress-to-egress extensions, and (4) non-critical ingress-to-egress extensions. Frames that violate this paragraph are erroneous, will produce unspecified results, and MAY be discarded. "MAY" is chosen to minimize the format-checking burden on transit RBridges.

If any extensions are present, those extensions, both flag and TLV, MUST be correctly summarized into the CHbH, CItE, and TLV summary bits.

### 2.4 Conflict of Extensions

It is possible for extensions to conflict. Two or more extensions can be present in a frame that direct an RBridge processing the frame to do conflicting things or to change its interpretation of later parts of the frame in conflicting ways. Such conflicts are resolved by applying the following rules in the order given:

1. Any frame containing extensions that require mutually incompatible changes in way later parts of the frame, after the extensions area, are interpreted or structured MUST be discarded. (Such extensions will be critical extensions, normally hop-by-hop critical extensions.)
2. Critical extensions override non-critical extensions.
2. Within each of the two categories of critical and non-critical extensions, the extension appearing first in lexical order in the frame always overrides an extension appearing later in the frame. Thus a conflict between an extended flag and a TLV extension is always resolved in favor of the extended flag. Extended flags with lower bit numbers are considered to have occurred before extended flags with higher bit numbers.

### 3. Specific Extended Header Flag

The table below shows the state of TRILL Header extended flag assignments and the location of the special Flow ID field. See Section 6 for IANA Considerations.

Bits	Purpose	Section
0-1	Critical Summary Bits	2.3
2	More extended flags	2.
3-4	available for critical hop-by-hop flags	
5	Alert Extended Flag	3.1
6-7	ECN	3.2
8-10	available for critical ingress-to-egress flags	
11-12	available for non-critical ingress-to-egress flags	
13-15	TLV Summary Bits	2.3.1.4
16-31	Flow ID	
32-39	available for critical hop-by-hop flags	
40-47	available for non-critical hop-by-hop flags	
48-55	available for critical ingress-to-egress flags	
56-63	available for non-critical ingress-to-egress flags	

Table 1. Extended Flag Extensions

#### 3.1 The Alert Extended Flag

The Alert Extended Flag indicates that the frame should be examined by the slow path at each hop. This is intended to alert transit R Bridges that implement this extension and to assist in the implementation of features such as a record route message.

#### 3.2 The ECN Extension

R Bridges MAY implement an ECN (Explicit Congestion Notification) extension [RFC3168]. If implemented, it SHOULD be enabled by default but can be disabled on a per R Bridge basis by configuration.

R Bridges that do not implement this extension or on which it is disabled simply (1) set bits 6 and 7 of the extended flags area to zero when they add an extensions area to a TRILL Header and (2) transparently copy those bits, if an extensions area is present, when they forward a frame with a TRILL Header.

An R Bridge that implements the ECN extension does the following, which correspond to the recommended provisions of [RFC6040], when that extension is enabled:

- o When ingressing an IP frame that is ECN enabled (non-zero ECN field), it MUST add an extensions area to the TRILL Header and copy the two ECN bits from the IP header into extended header flags 6 and 7.
- o When ingressing a frame for a non-IP protocol, where that protocol has a means of indicating ECN that is understood by the RBridge, it MAY add an extensions area to the TRILL Header with the ECN bits set from the ingressed frame.
- o When forwarding a frame encountering congestion at an RBridge, if an extensions area is present with extended flags 6 and 7 indicating ECN-capable transport, the RBridge MUST modify them to the congestion experienced value.
- o When egressing an IP frame, the RBridge MUST set the outgoing native IP frame ECN field to the codepoint at the intersection of the values for that field in the encapsulated IP frame (row) and the TRILL extended Header ECN field (column) in Table 3 below or drop the frame in the case where the TRILL header indicates congestion experienced but the encapsulated native IP frame indicates a not ECN-capable transport. (Such frame dropping is necessary because IP transport that is not ECN-capable requires dropped frames to sense congestion.)
- o When egressing a non-IP protocol frame with a means of indicating ECN that is understood by the RBridge, it MAY set the ECN information in the egressed native frame by combining that information in the TRILL extended header and the encapsulated non-IP native frame as specified in Table 3.

The following table is modified from [RFC3168] and shows the meaning of bit values in TRILL Header extended flags 6 and 7, bits 6 and 7 in the IPv4 TOS Byte, and bits 6 and 7 in the IPv6 Traffic Class Octet:

Binary	Meaning
-----	-----
00	Not-ECT (Not ECN-Capable Transport)
01	ECT(1) (ECN-Capable Transport(1))
10	ECT(0) (ECN-Capable Transport(0))
11	CE (Congestion Experienced)

Table 2. ECN Field Bit Combinations

Table 3 below (adapted from [RFC6040]) shows how, at egress, to combine the ECN information in the extended TRILL Header ECN field with the ECN information in an encapsulated frame to produce the ECN information to be carried in the resulting native frame.

Inner Native Header	Arriving TRILL Header ECN Field			
	Not-ECT	ECT(0)	ECT(1)	CE
Not-ECT	Not-ECT	Not-ECT(*)	Not-ECT(*)	<drop>(*)
ECT(0)	ECT(0)	ECT(0)	ECT(1)	CE
ECT(1)	ECT(1)	ECT(1)(*)	ECT(1)	CE
CE	CE	CE	CE(*)	CE

Table 3: Egress ECN Behavior

An RBridge detects congestion either by monitoring its own queue depths or from participation in a link-specific protocol. An RBridge implementing the ECN extension MAY be configured to add congestion experienced marking using ECN to any frame with a TRILL Header that encounters congestion even if the frame was not previously marked as ECN-capable or did not have an extensions area.

#### 4. Specific TLV Extension

The table below shows the state of TRILL Header TLV extension Type assignment. See Section 6 for IANA Considerations.

Type	Purpose	Section
0x00	reserved	
0x00-0x7F	available	
0x80	Test/Pad	4.1
0x81-0xFE	available	
0xFF	reserved	

Table 4. TLV Extension Types

The following subsection specifies a particular TRILL TLV extension.

##### 4.1 Test/Pad Extension

This extension is intended for testing and padding.

A specific meaning for this extension with the critical flag set will not be defined so, in that form, it MUST always be treated as an unknown critical extension. If the critical flag is not set, the extension does nothing. In either case, it may be any length that will fit. Thus, for example, in the non-critical form, it can be used to cause the encapsulated frame staring right after the extensions area to be 64-bit aligned or for testing purposes.

- o Type is 0x80.
- o Length is variable. The value is ignored.
- o IE may be zero or one. This extension has both hop-by-hop and ingress-to-egress versions.
- o NC is zero for the pad extension and one for the test extension.
  - + The non-critical version of this extension does nothing.
  - + The critical version of this extension MUST always be treated as an unknown critical extension.
- o MU may be zero or one except that it must be zero if the other flags indicate the extensions is a critical hop-by-hop extension. This extension may be flagged as mutable or immutable.

## 5. Additions to IS-IS

RBridges use IS-IS PDUs to inform other RBridges which extensions they support. The specific IS-IS PDUs, TLVs, or sub-TLVs used to encode and advertise this information are specified in a separate document. Support for critical extensions **MUST** be advertised. Support for non-critical extensions **MAY** be advertised unless the specification of a particular non-critical extension imposes a requirement higher than "MAY" for the advertising of that extension by RBridges that implement it.

## 6. IANA Considerations

IANA will create two subregistries within the TRILL registry. A "TRILL Extended Header Flags" subregistry that is initially populated as specified in Table 1 in Section 3. And a "TRILL TLV Extension Types" subregistry that is initially populated as specified in Table 4 in Section 4. References in both of those tables to sections of this document are to be replaced in the IANA subregistries by references to this document as an RFC.

New TRILL bit extensions and TLV extension types are allocated by IETF Review [RFC5226].

## 7. Security Considerations

For general TRILL protocol security considerations, see [RFCtrill].

In order to facilitate authentication, extensions SHOULD be specified so they do not have alternative equivalent forms. Authentication of anything with alternative equivalent forms almost always requires canonicalization that an authenticating RBridge ignorant of the extension would be unable to do and that may be complex and error prone even for an RBridge knowledgeable of the extension. It is best for any extension to have a unique encoding.

## 8. Acknowledgements

The following are thanked for their contributions: Bob Briscoe.



## 9. References

Normative and informative references for this document are given below.

### 9.1 Normative References

- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3168] - Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.
- [RFC5226] - Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC6040] - Briscoe, B., "Tunneling of Explicit Congestion Notification", RFC 6040, November 2010
- [RFCtrill] - Perlman, R., D. Eastlake, D. Dutt, S. Gai, and A. Ghanwani, "RBridges: Base Protocol Specification", draft-ietf-trill-rbridge-protocol-16.txt, in RFC Editor's queue.

### 9.2 Informative References

None.

## Change History

The sections below summarize changes between successive versions of this draft. RFC Editor: Please delete this section before publication.

## Version 00 to 02

Change the requirement for TLV option ordering to be strictly ordered by the value of the top nine bits of their first two bytes so that the MU bit is included.

Specify meaning of mutability bit for hop-by-hop options.

Fix length of Flow ID Value at 2.

Require that options that may significantly affect the interpretation or format of subsequent parts of the frame be critical options.

## Version 02 to 03

Move Test/Pad extension into this document from the More Options draft and move the More Flags option from this document into the More Options draft.

Prohibit multiple occurrences of a TLV option in a frame.

## Version 03 to 04

Restructure the bit encoded options area so that the initial 32 bits include a 16 bit Flow ID, various TLV-option-present bits, and a more extended flags bit that means another 32 bits of extended flags are present.

Change the Length of TLV encoded options so that it is in units of 4 bytes, not 1, resulting in a bigger Type field.

Update Explicit Congestion Notification to follow RFC 6040.

Rename "bit encoded options" to be "extended header flags" or "extended flags".

Version 04 to 05

Generally replace "option" with "extension".

Add the Alert critical hop-by-hop flag extension.

Replace MT with MU to avoid possible confusion with multiple topologies.

Authors' Addresses

Donald Eastlake  
Huawei Technologies  
155 Beaver Street  
Milford, MA 01757

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

Anoop Ghanwani  
Brocade Communications Systems  
130 Holger Way  
San Jose, CA 95134 USA

Phone: +1-408-333-7149  
Email: anoop@brocade.com

Vishwas Manral  
IP Infusion Inc.  
1188 E. Arques Ave.  
Sunnyvale, CA 94089 USA

Tel: +1-408-400-1900  
email: vishwas@ipinfusion.com

Caitlin Bestler  
Quantum  
1650 Technology Drive , Suite 700  
San Jose, CA 95110

Phone: +1-408-944-4000  
email: cait@asomi.com

## Copyright and IPR Provisions

Copyright (c) 2011 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  
Expires: December 2011

David Melman  
Tal Mizrahi  
Marvell  
Donald Eastlake  
Huawei  
June 30, 2011

FCoE over TRILL  
draft-mme-trill-fcoe-00.txt

## Abstract

Fibre Channel over Ethernet (FCoE) and TRILL are two emerging standards in the data center environment. While these two protocols are seemingly unrelated, they have a very similar behavior in the forwarding plane, as both perform hop-by-hop forwarding over Ethernet, modifying the packet's MAC addresses at each hop. This document describes an architecture for the integrated deployment of these two protocols.

## 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/lid-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 December 30, 2011.

## Copyright Notice

Copyright (c) 2011 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.....	2
2. Abbreviations.....	3
3. FCoE over TRILL.....	4
3.1. FCoE over a TRILL Cloud.....	4
3.2. FCoE over RBridge.....	5
3.2.1. FCRB.....	5
3.2.2. Topology.....	7
3.2.3. The FCRB Flow.....	8
4. Security Considerations.....	10
5. IANA Considerations.....	10
6. Acknowledgments.....	11
7. References.....	11
7.1. Normative References.....	11
7.2. Informative References.....	11

## 1. Introduction

Data center networks are rapidly evolving towards a consolidated approach, where Ethernet is used as the common infrastructure for all types of traffic. Storage traffic, which was traditionally dominated by the Fibre Channel (FC) protocol suite, is evolving towards Fibre Channel over Ethernet (FCoE), where native FC packets are encapsulated with an FCoE encapsulation over an Ethernet header.

Traffic between two FCoE end nodes (ENodes) is forwarded through one or more FCoE Forwarders (FCF). An FCF takes a forwarding decision based on the Fibre Channel destination ID (D\_ID), and enforces security policies between ENodes, also known as zoning. Once an FCF takes a forwarding decision, it modifies the source and destination MAC addresses of the packet, to reflect the path to the next hop FCF or ENode. FCFs use a routing protocol called Fabric Shortest Path First (FSPF) to find the optimal path to each destination. An FCF typically has one or more native Fibre Channel interfaces, allowing



it to communicate with native Fibre Channel devices, e.g., storage arrays.

TRILL [RFC6234] is a protocol for transparent least cost routing, where RBridges forward traffic to their destination based on a least cost route, using a TRILL encapsulation header. RBridges forward TRILL-encapsulated packets based on the Egress RBridge Nickname in the TRILL header. An RBridge forwards a TRILL-encapsulated packet after modifying its MAC addresses to reflect the path to the next-hop RBridge, and decrementing a Hop Count field.

TRILL and FCoE bear a strong resemblance in their forwarding planes. Both protocols take a forwarding decision based on protocol addresses above Layer 2, and modify the Ethernet MAC addresses on a per-hop basis. Each of the protocols uses its own routing protocol rather than using any type of bridging protocol such as spanning tree protocol [802.1Q] or the Shortest Path Bridging protocol [802.1aq].

FCoE and TRILL are both targeted at the data center environment, and their concurrent deployment is self-evident. This document describes an architecture for the integrated deployment of these two protocols.

## 2. Abbreviations

ENode     FCoE Node such as server or storage array

EoR       End of Row

FC        Fibre Channel

FCF       Fibre Channel Forwarder

FCoE      Fibre Channel over Ethernet

FCRB      Fibre Channel forwarder over RBridge

FDF       Fibre Channel data-plane Forwarder

FSPF      Fabric Shortest Path First

LAN       Local Area Network

RBridge   Routing Bridge

SAN       Storage Area Network

ToR       Top of Rack

TRILL    Transparent Interconnection of Lots of Links

WAN      Wide Area Network

### 3. FCoE over TRILL

#### 3.1. FCoE over a TRILL Cloud

The simplest approach for running FCoE traffic over a TRILL network is presented in Figure 1. The figure illustrates a TRILL-enabled network, where FCoE traffic is transparently forwarded over the TRILL cloud. The figure illustrates two ENodes, a Server and an FCoE Storage Array, an FCF, and a native Fibre Channel SAN connected to the FCF.

FCoE traffic between the two ENodes is sent from the first ENode over the TRILL cloud to the FCF, and then back through the TRILL cloud to the second ENode.

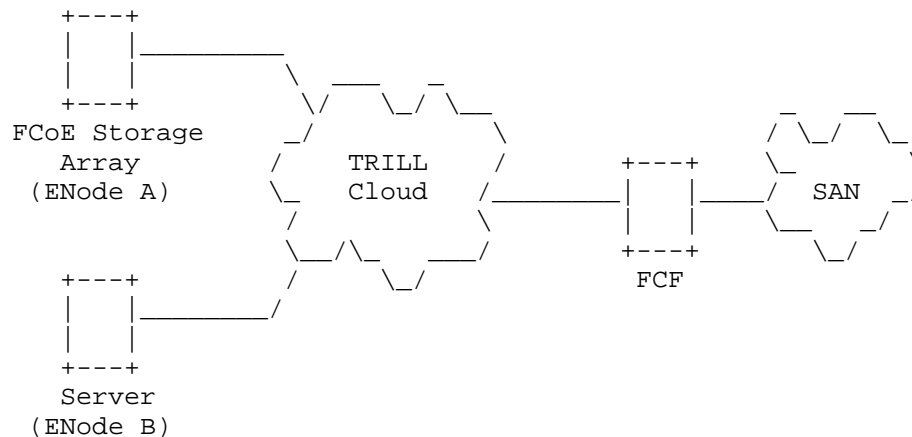


Figure 1 The "Separate Cloud" Approach

The configuration in Figure 1 separates the TRILL cloud(s) and the FCoE cloud(s). The TRILL cloud forwards FCoE traffic as standard Ethernet traffic, and appears to the ENodes and FCF as an Ethernet LAN.

The main drawback of the Separate Cloud approach is that RBridges and FCFs are separate nodes in the network, resulting in more cabling and boxes, and communication between ENodes usually requires two TRILL cloud traversals with twice as many hops. As mentioned above,

data center networking is converging towards a consolidated and cost effective approach, where the same infrastructure and equipment is used for both data and storage traffic, and where high efficiency and minimal number of hops are important factors when designing the network topology.

### 3.2. FCoE over RBridge

#### 3.2.1. FCRB

Rather than the Separate Cloud approach discussed in the previous subsection, an alternate approach is presented, where each switch incorporates both an FCF entity and an RBridge entity. This consolidated entity is referred to as FCoE-forwarder-over-RBridge (FCRB).

Figure 2 illustrates an FCRB, and its main building blocks. An FCRB can be functionally viewed as two independent entities:

- o An FCoE Forwarder (FCF) entity or an FCoE data-plane Forwarder (FDF) entity.
- o An RBridge entity.

The FCF/FDF entity is connected to one of the ports of the RBridge, and appears to the RBridge as a native Ethernet host. A detailed description of the interaction between the layers is presented in Section 3.2.3.

An FCF and an FDF are similar in the data-plane, and differ in the control-plane. Namely, FCF performs the session initiation with the ENode, while an FDF does not take part in this procedure.

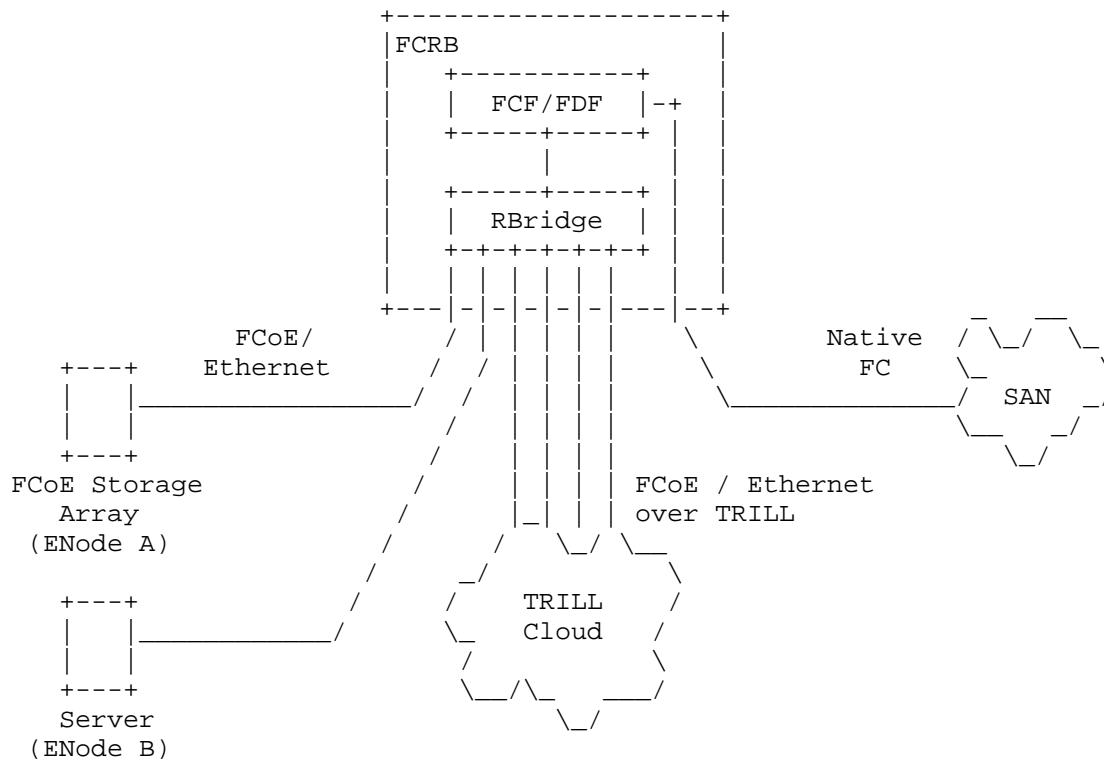


Figure 2 FCRB Entity in the Network

The FCRB entity maintains layer independence between the TRILL and FCoE protocols, while enabling both protocols on the same network.

It is noted that FCoE traffic is always forwarded through an FCF/FDF, and cannot be forwarded directly between two ENodes. Thus, FCoE traffic between ENodes A and B in the topology in Figure 1 is forwarded through the path

ENode A-->TRILL cloud-->FCF-->TRILL cloud-->ENode B

Traffic between A and B in the topology in Figure 2 is forwarded through the path

ENode A-->FCRB-->ENode B

Hence, the usage of FCRB entities allows TRILL and FCoE to use common infrastructure and equipment, as opposed to the Separate Cloud topology presented in Figure 1.

### 3.2.2. Topology

The network configuration illustrated in Figure 3 shows a typical topology of a data center network. Servers are hierarchically connected through Top-of-Rack (ToR) switches, and End-of-Row (EoR) racks. The EoR switches to other clouds, such as an external WAN, or a native FC SAN.

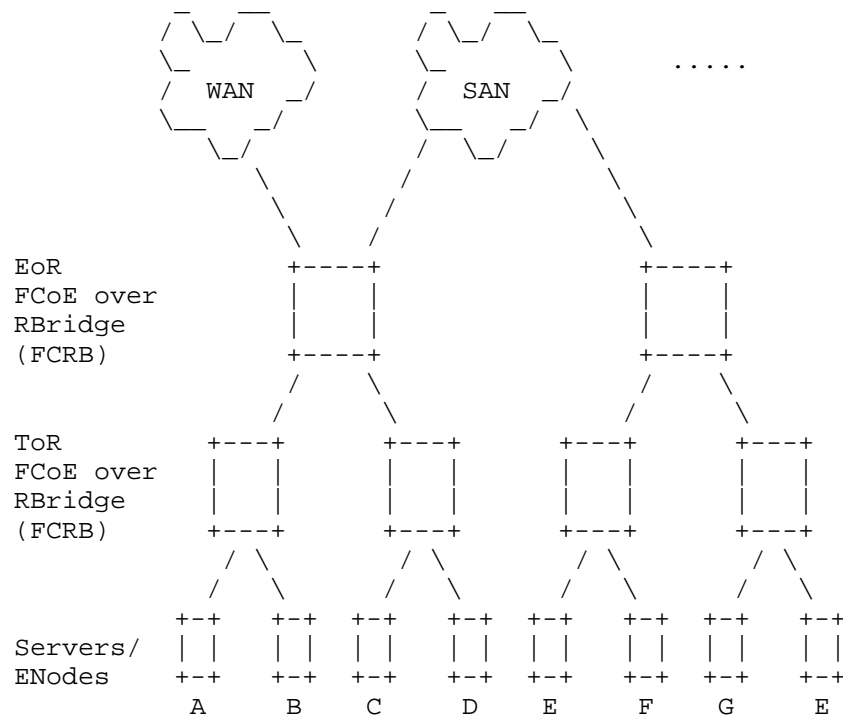


Figure 3 FCoE over RBridge Topology

### 3.2.3. The FCRB Flow

FCoE traffic sent between two ENodes, A and B, is transmitted through the ToR FCRB, since A and B are connected to the same ToR. Traffic between A and C must be forwarded through the EoR FCRB.

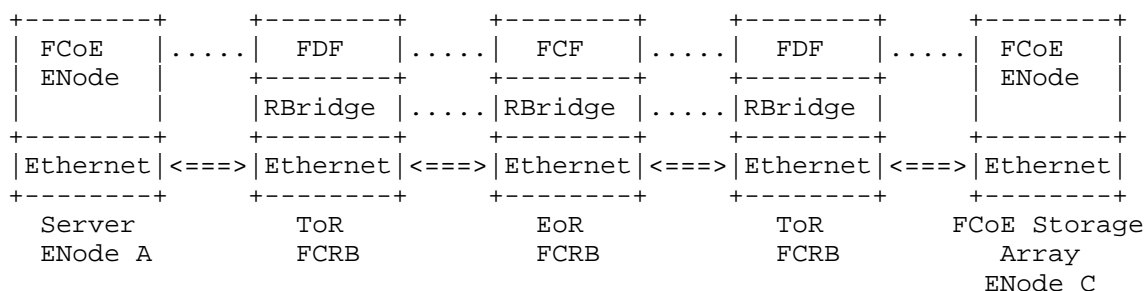


Figure 4 Traffic between two ENodes - Example

Figure 4 illustrates the traffic between ENodes A and C that are not connected to the same ToR.

- o FCoE traffic from A is sent to the ToR over the Ethernet interface.
- o The RBridge entity at the ToR forwards the packet to the FDF, analogous to forwarding between two Ethernet hosts. The FDF entity at the ToR takes a forwarding decision, and updates the destination MAC address of the packet to the address of the EoR FCF. The packet is then forwarded to the RBridge entity, where it is encapsulated in a TRILL header, and sent to the EoR FCRB.
- o The RBridge entity in the EoR FCRB, acting as the egress RBridge, decapsulates the TRILL header and forwards the FCoE packet to the FCF entity. The FCF takes a forwarding decision and updates the MAC address of the packet according to the next hop ToR. The packet is then forwarded to the RBridge and encapsulated with a new TRILL header.
- o At the ToR FCRB, the packet reaches the final egress RBridge, and the TRILL encapsulation is removed. The FDF then forwards the packet to the RBridge entity after updating its MAC addresses. The RBridge entity forwards the packet to the target ENode.

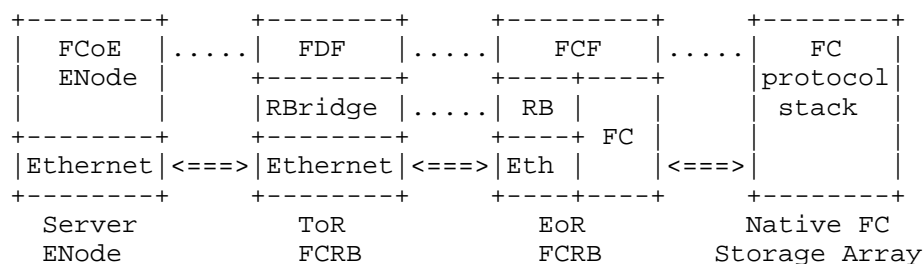


Figure 5 Example Traffic between ENode & Native FC Storage Array

Figure 5 illustrates traffic sent between an ENode and an FC Storage Array, following the network topology in Figure 3.

- o FCoE traffic from the ENode is sent to the ToR over the Ethernet interface.
- o The RBridge entity at the ToR forwards the packet to the FDF. The FDF entity at the ToR takes a forwarding decision and updates the destination MAC address of the packet to the address of the EoR FCF. The packet is then forwarded to the RBridge entity, where it is encapsulated in a TRILL header, and sent to the EoR FCRB.
- o The egress RBridge entity at the EoR FCRB decapsulates the TRILL header, and forwards the FCoE packet to the FCF entity. The packet is then forwarded as a native FC packet through the FC interface to the native FC node.

#### 4. Security Considerations

For general TRILL Security Considerations see [RFCtrill].

For general FCoE Security Consideration see Annex D of [FC-BB-5].

There are no additional security implications imposed by this document.

#### 5. IANA Considerations

There are no IANA actions required by this document.

RFC Editor: please delete this section before publication.



## 6. Acknowledgments

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

## 7. References

### 7.1. Normative References

- [RFCTRILL] Perlman, R., Eastlake, D., Dutt, D., Gai, S  
.,  
Ghanwani, A., "RBridges: Base Protocol Specification",  
draft-ietf-trill-rbridge-protocol (work in progress),  
March 2010.

### 7.2. Informative References

- [FC-BB-5] ANSI INCITS 462: Information Technology - Fibre  
Channel - Backbone - 5 (FC-BB-5).
- [FC-BB-6] "FIBRE CHANNEL - BACKBONE - 6 (FC-BB-6)", work in  
progress, Rev 1.02, Oct 2010.
- [802.1Q] "IEEE Standard for Local and metropolitan area networks  
- Virtual Bridged Local Area Networks", IEEE Std  
802.1Q-2011, May 2011.
- [802.1aq] "IEEE Standard for Local and metropolitan area  
networks - Shortest Path Bridging", work in progress,  
June 2011.

## Authors' Addresses

David Melman  
Marvell  
6 Hamada St.  
Yokneam, 20692 Israel  
  
Email: davidme@marvell.com

Tal Mizrahi  
Marvell  
6 Hamada St.  
Yokneam, 20692 Israel  
  
Email: talmi@marvell.com

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

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



INTERNET-DRAFT  
Intended Status: Proposed Standard  
Expires: January 12, 2012

Mingui Zhang  
Dacheng Zhang  
Huawei  
July 11, 2011

Adaptive VLAN Assignment for Data Center RBridges  
draft-zhang-trill-vlan-assign-01.txt

Abstract

If RBridges are casually assigned as Appointed Forwarders for VLANs without considering the number of MAC addresses and traffic load of these VLANs, it may overload some of the RBridges while leave other RBridges lightly loaded, which reduces the scalability of a RBridge network and undermines its performance.

There can only be a single appointed forwarder for one VLAN carried by a LAN link at the same time, even if this LAN link is attached to multiple points of an RBridge campus. This limitation not only wastes the available access bandwidth of RBridge campus but also reduces its reliability.

A new protocol is designed in this document to support the adaptive VLAN assignment (or Appointed Forwarder selection) based on the forwarders' reporting of their usage of MAC tables and available bandwidth. Link aggregation is proposed to overcome the single forwarder limitation of TRILL.

Status of this Memo

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

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

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

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

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

## Copyright and License Notice

Copyright (c) 2011 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 . . . . .	4
2. Data Center RBridge . . . . .	4
2.1. Scalability . . . . .	4
2.2. East-West Capacity Increase . . . . .	4
2.3. Virtualization . . . . .	5
3. MAC Entries Balancing among VLANs . . . . .	5
4. Load Balancing among VLANs . . . . .	7
4.1. Egress Traffic . . . . .	7
4.2. Ingress Traffic . . . . .	8
5. Load Balancing within a VLAN . . . . .	8
6. Definition of sub-TLVs . . . . .	10
6.1. MAC Entries Report sub-TLV . . . . .	11
6.2. Traffic Bit Rate Report sub-TLV . . . . .	12
7. Security Considerations . . . . .	14
8. IANA Considerations . . . . .	15
9. References . . . . .	15
9.1. Normative References . . . . .	15
9.2. Informative References . . . . .	15
Author's Addresses . . . . .	17

## 1. Introduction

The scales of Data Center Networks (DCNs) are expanding very fast these years. In DCNs, Ethernet switches and bridges are abundantly used for the interconnection of servers. The plug-and-play feature and the simple management and configuration of Ethernet are appealing to the DCN providers. A whole DCN can be a simple large layer 2 Ethernet which is either built on a real network or on a virtualization platform.

Cloud Computing is growing up from DCNs which can be seen as a virtualization platform that provides the reuse of the network resources of DCNs. A lot of cloud applications have been developed by DCN providers, such as Amazon's Elastic Compute Cloud (EC2), Akamai's Application Delivery Network (ADN) and Microsoft's Azure. Cloud Computing clearly brings new challenges to the traditional Ethernet. The scales of the DCNs are becoming too large to be carried on the traditional Ethernet. The valuable MAC-tables of the bridges are running out of use for storing millions of MAC addresses. The broadcast of ARP messages consumes too much bandwidth and computing resources. The mobility of end stations brings dynamics to the network which can be a heavy burden if the management and configuration of the network involves too much manpower. The Spanning Tree Protocol used in the traditional Ethernet is outdated since there is only a single viable path on the tree for a node pair and this path is not always the best path (e.g., shortest path).

RBridges are designed to improve the shortcomings of the traditional Ethernet. To make use of the rich connections, RBridges introduce multi-pathing to the Ethernet to break the single-path constraint of STP. Multiple points of attachment is a basic feature supported by RBridges and common for Data Center Bridges. This feature not only increases the "east-west" capacity but also greatly enhances the reliability of DCNs [VL2] [SAN]. If several RBridges are attached to a bridged LAN link at the same time, the DRB is responsible for the assignment of a VLAN to one of the RBridges as the Appointed forwarder. However, the current VLAN assignment is done in an one-way manner. The DRB casually assign a VLAN to an RBridge attached to the local link without knowing its available MAC-table entries or bandwidth. The appointed forwarder does not feedback the utilization of its MAC-table or bandwidth either.

This document proposes the solutions for balancing the load among VLANs and the load within a single VLAN. Two types of sub-TLVs are defined, with which a forwarder can report its MAC entries and traffic bit rate respectively. By gathering these report messages, the VLAN assignment can be done in a way that the usage of the MAC tables and bandwidth of the attached RBridges are balanced among the

VLANs.

### 1.1. Terminology

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

## 2. Data Center RBridge

Data Center Networks grow rapidly recently. Ethernet is widely used in data centers because of its simple management and plug-and-play features. However, there are shortcomings of Ethernet. RBridges are designed to improve these shortcomings. In this section, we analyze the characteristics of the DCNs that impact the design of RBridges and reveal why the adaptive VLAN assignment is important for RBridges to be used in DCNs.

### 2.1. Scalability

In the past years, a large DCN is typically composed of tens of thousands servers interconnected through switches and bridges. In the future cloud computing era, there can be as many as millions of servers in one DCN. The management of the numerous MAC addresses of the servers on the layer2 devices will become more and more complex. RBridges are aimed to replace the traditional bridges. The valuable CAM-tables on RBridges can easily be used up if they are not used reasonably [CAMtable]. For RBridges to be widely used in DCNs, the VLANs should be assigned to the RBridges in a manner that the MAC entries of the VLANs on the RBridges are balanced.

### 2.2. East-West Capacity Increase

The Spanning Tree Protocol (STP) in the traditional LAN blocks some ports of the bridges for the purpose of loop avoidance. However, the side-effects of STP are obvious. The link bandwidth attached to the blocked ports are not used which greatly wastes the capacity of the network. On the tree topology, the communication between the bridges of the left branch and right branch must transit the single root bridge, which forms a "hair-pin turn".

With the rapid increase of the amount of servers in DCNs and their traffic demand, it is urgent to break the constraint of STP and enhance the "east-west" capacity of DCNs which are always richly connected. RBridges use the multi-path routing to set up the data plane of a TRILL network. Multiple RBridges may be attached to the same LAN link, which offers multiple access points to the LAN link. The hosts on this LAN link is therefore multi-homed to a TRILL network. All the attached RBridges can act as the packet forwarder for the VLANs carried on this LAN link. In the worst case, all the VLANs are probably assigned to a single RBridge. Under this scenario, the ingress capacity on the other RBridges is wasted. It is necessary to balance the traffic load of the VLANs among these RBridges through the assignment of the VLANs.

### 2.3. Virtualization

Virtualization is important for increasing the utilization of network resources in DCNs. For example, the VPNs can be used to separate the traffic from different services therefore they can be carried on the same pool of resources. When the VPNs is carried over a TRILL network, RBridges can use a VLAN tag to identify each VPN. However, the use of VLANs multiplies the entries in the MAC table of the RBridges. Since a host can be a member of several VLANs at the same time, the RBridges have to store multiple copies of its MAC address in its precious MAC table.

Virtual Machines (VM) are widely used in DCNs. A physical host can support multiple VMs and each of the VMs has to be identified by one MAC address that is need to be stored in the MAC tables of the RBridges. This seriously increases the numbers of MAC entries in RBridges. Moreover, the number of VMs in a VLAN is not necessarily equal to the number of the physical hosts. VMs are spawned or destroyed based on the demand of the applications. They can also migrate from one location to another, which may be either an in-service or out-of-service move. VMs bring about the volatility of the size of VLANs. It is hard for a TRILL network to provide one static VLAN assignment based on the numbers of physical hosts of VLANs that is proper for all applications all the time. It is necessarily to do VLAN assignment adaptively.

### 3. MAC Entries Balancing among VLANs

A CAM-table on a switch is expensive, which is a major constraint on the scalability of Ethernet [CAMtable]. When a RBridge is used to connect lots of hosts in large Data Center Networks, the entries of the CAM-table can easily be used up. The network should be tactically interconnected and the valuable MAC table entries should be used economically.



RBridges support multiple points of attachment [TRILLbase]. When RBridges are used in a DCN to form a TRILL network, a LAN link MAY have multiple access points to this network. All the access RBridges are able to act as the packet forwarder of the VLANs carried on this LAN link. The DRB of this LAN link is responsible to pick out one of the RBridge attached to this LAN link as the appointed forwarder for each VLAN-x. In other words, the DRB assigns VLAN-x to one of the RBridge. For an assigned VLAN, its forwarder is not only responsible for forwarding the packets but also need to store the active MAC addresses of the hosts on this VLAN.

If the VLANs on the LAN link are not appointed properly, some of the RBridges's MAC tables are easily to be used up while the other RBridges are left idle. Take Figure 2.1 as an example, there are four VLANs carried on the LAN link: w, x, y and z. There are two hosts in both VLAN-w and VLAN-x and one host in both VLAN-y and VLAN-z. RB1 and RB2 are both attached to this LAN link. RB1 is elected as the Designated RBridge who is responsible to choose the appointed forwarder for the above VLANs. The figure shows that VLAN-w,x are assigned to RB1 and VLAN-y,z are assigned to RB2. Obviously, this assignment is not balanced, since the MAC table of RB1 has four entries while the MAC table of RB2 only has two entries. If the DRB can reassign VLAN-w to RB2 and reassign VLAN-y to RB1, both RBridges will have three MAC entries, therefore a more balanced assignment is achieved.

In order to assign the VLANs in a balanced way, the DRB need to know the usage of the MAC tables of its appointed forwarders and the sum of the MAC addresses in each VLAN. Since the RBridges only store the active MAC addresses and a virtual machine can move from one location to another, the MAC entries a VLAN occupy on an RBridge varies from time to time. The assignment of the VLANs cannot be done once for all. It is necessary for the DRB to do the assignment adaptively taking the usage of MAC tables of its appointed forwarders into consideration. Therefore, in Section 5.1, the MAC Entries Report sub-TLV is defined to deliver this kind of information from a forwarder to a DRB.

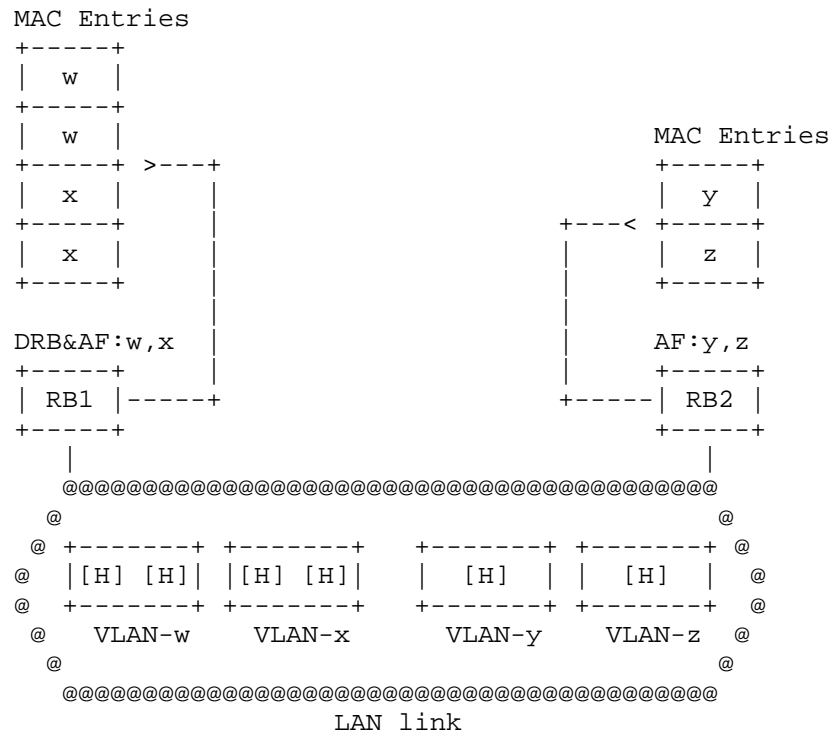


Figure 2.1: Unbalanced assignment among VLANs

#### 4. Load Balancing among VLANs

The traffic from the TRILL network to the local LAN link is called egress traffic while the traffic from the local LAN link to the TRILL network is called ingress traffic. A forwarder RBridge acts as both the ingress and egress point of a VLAN's traffic. The assignment of the appointed forwarder for each VLAN affects both the egress and ingress traffic load distribution.

##### 4.1. Egress Traffic

One RBridge MAY have multiple ports attached to the same local LAN link. These ports are called "port group" [TRILLbase]. When a DRB assigns a VLAN to an RBridge, its total available egress bandwidth of the port group needs to be taken into consideration. Using the TLV defined in Section 5.2, the load of the egress points are reported from the appointed forwarders to the DRB on the LAN link. The assignment SHOULD NOT cause congestion to an already busy egress point.

After VLAN-x has been assigned to an RBridge, the forwarding port assignment of one of the port group to VLAN-x as the forwarding port is entirely a local matter. Since a LAN link is a STP domain, more than one forwarding port for one VLAN will cause a loop. The forwarder MUST assign one and only one port for each VLAN. Load balancing can be realized through splitting the load among different VLANs as suggested in Section 4.4.4 of [TRILLbase].

#### 4.2. Ingress Traffic

After the known unicast packets enter the TRILL network from the ingress RBridge, they can be sent through the paths starting at this ingress point. Since the DRB knows the whole topology of the TRILL network, it can figure out these paths as well. Therefore, the DRB should take the available bandwidth of these paths into consideration when assigning the appointed forwarder of a VLAN. Any assignment that is possible to congest an already busy ingress point or a path should be avoided.

Traffic Matrices are usually taken as the input to the traffic engineering methods [TE]. The work in this section is actually changing the Traffic Matrices of the TRILL network. If traffic engineering is used in TRILL networks, the forwarder appointment mechanism should work together with the traffic engineering method to in order to achieve a more balanced global traffic distribution of the whole network. The DRB can also collect the probing messages used in the traffic engineering and then assign the VLAN according to the bandwidth utilization. However, the design of this kind of cooperative mechanism for balancing the ingress traffic is left as future work when traffic engineering solutions are begin to be used on TRILL networks [TBD].

#### 5. Load Balancing within a VLAN

Section 4 addresses the load balancing among different VLANs, while this section talks about the load balancing with finer granularity: balancing the load for a single VLAN.

For loop avoidance, there can ONLY be a single appointed forwarder ingressing and egressing native frames on a link for a specific VLAN-x at the same time [TRILL-AF]. Take Figure 2.1 as an example, although RB1 and RB2 both can perform frame forwarding for VLAN-X, DRB can only appoint one of them to be the appointed forwarder, the other one will be inhabited from ingressing and egressing frames of VLAN-x. This single forwarder mechanism does not take the full advantage of the multiple attachment character of TRILL networks which not only wastes the available access bandwidth but also reduces the network resilience.

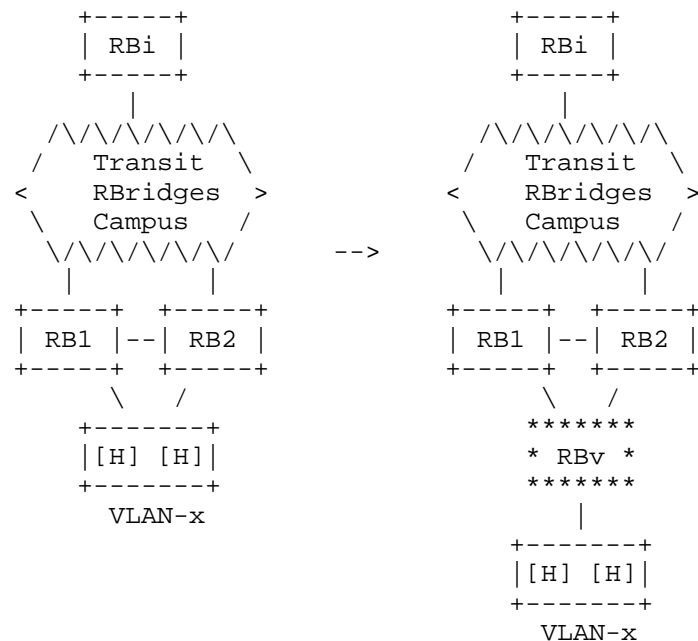


Figure 5.1: Sample network for the illustration of Link Aggregation

Link aggregation is widely used to address two problems with Ethernet connections: bandwidth limitations and lack of resilience [802-1AX] [vPC+] [Brocade]. R Bridges can make use of link aggregation to break the above limitation of TRILL networks. Figure 5.1 gives a sample network for the illustration of link aggregation. RB1 and RB2 are both attached to the local link which carries VLAN-x. They are bundled together as a virtual R Bridge (RBv) to act as VLAN-x's forwarder. The nickname of RBv will be configured and known to RB1 and RB2. The virtual links between RB1 and RBv and RB2 and RBv will be announced in the LSPs of RB1 and RB2. Other R Bridges will believe there really is a node RBv connecting RB1 and RB2 on the campus. When ingress and egress native frames, the bundled R Bridges should act as follows.

1. Unicast frames: For a native frame comes from the local link, the traffic is balanced across the access links based on hash values computed based on fields of the packet header. The selected RBridge by the hashing function will forward the native frame on behalf of the virtual RBridge and fill in the nickname of RBv instead of its own nickname. For the TRILL data frames destined to the local link, the receiver RBridge (a member of the bundling group) directly deliver the packet onto the local link. Remote R Bridges use SPF to compute paths to the virtual RBridge. Since

Equal Cost Multi-Pathing is supported by RBridges and multi-topology is to be supported by RBridges as well, it is possible that both of the members of the bundling group are selected as the transit RBridges. Load balance can be achieved through multi-pathing.

2. Multi-destination frames (as defined in [TRILLbase]): For the ingressing and egressing of multi-destination frames, another hashing function is used to assign each of the frame to a member of the bundling group. For example, the source MAC address and the ingress fields of the native frame can be used as the seed of this hashing function. If the receiver RBridge finds that the frame is assigned to the peer rather than itself, it will discard the frame since its peer will deliver the frame. Otherwise, if the frame is right assigned to the receiver RBridge, it will deliver this frame. In particular, for an unknown unicast frame which should be broadcast by the peer RBridge, the frame will be transmitted to the peer as a unicast frame. The peer will broadcast this unknown unicast frame.
3. MAC synchronization: The MAC addresses SHOULD be synchronized between the bundling members through ESADI immediately after they are learned from the data plane. A MAC address learned through ESADI from the peer is stored as if it is locally learned. Afterwards, a frame to this MAC address can be delivered to the local link by either of the bundling members.
4. Link failures: The links between the local link and RB1 and RB2 are protected by each other. For example, when RB1 detects that it is disconnected from the local link, it will transmit the known unicast frames to RB2 for further delivery. For a multi-destination frame or unknown unicast frame that should be delivered by RB1 according to the hashing function, a suggested option is to send the frame to RB2 through a reserved outer VLAN. RB2 will deliver this multi-destination frame without considering the hashing function.

In [TRILLbase], in order to suppress loops, multiple appointed forwarders for the same VLAN on a same local link is prohibited. This limitation should be relaxed under the link aggregation scenario. For example, the HELLO inhabitation rule defined in [TRILLbase] Section 4.2.4.3 should be removed for the bundling members.

#### 6. Definition of sub-TLVs

The Appointed Forwarders TLV has already been defined in [TRILLtlv]. With this TLV, the DRB can appoint an RBridge on the local link to be

the forwarder for each VLAN. However, there is no feedback from the appointed forwarder whether the assignment is reasonable. Two sub-TLVs are defined in this section to open the feedback passageway. They can be used by the appointed forwarder to report the number of MAC addresses and traffic load of VLANs in the reverse direction to the DRB. Through the collection of these report messages (these messages can be stored in the MIB of DRB [TRILLmib]), the DRB will have a vision of the MAC tables usage and bandwidth utilization of the R Bridges on the LAN link. Based on this vision, the DRB can have a adaptive VLAN assignment.

#### 6.1. MAC Entries Report sub-TLV

The appointed forwarder use MAC Entries Report sub-TLV to report the usage of its MAC table to the DRB. It has the following format:

```

+-----+
|Type=MACetrRep | (1 byte)
+-----+
| Length        | (1 byte)
+-----+
| DRB Nickname  | (2 bytes)
+-----+
| Maximum MAC Entries | (2 bytes)
+-----+
| Available MAC Entries | (2 bytes)
+-----+
| MAC Entries of VLAN (1) | (4 bytes)
+-----+
| ..... | (4 bytes)
+-----+
| MAC Entries of VLAN (N) | (4 bytes)
+-----+

```

where each MAC Entries of VLAN is of the form:

```

+-----+
| RESV | VLAN ID | (2 bytes)
+-----+
| The Number of MAC Entries | (2 bytes)
+-----+

```

- o Type: MAC Entries Report sub-TLV.
- o Length: 6+4n bytes, where n is the number of VLANs that the appointed forwarder selects to report their numbers of MAC entries in its MAC table.

- o DRB Nickname: The nickname of the Designated RBridge of the local link.
- o Maximum MAC Entries: The maximum number of the entries of the MAC table of the appointed forwarder.
- o Available MAC Entries: The number of available entries of the MAC table of the appointed forwarder.
- o RESV: 4 bits that MUST be sent as zero and ignored on receipt.
- o VLAN ID: This field identifies one of the VLANs that assigned to the appointed forwarder.
- o The Number of MAC Entries: The number of MAC Entries that the given VLAN occupies in the MAC table of the appointed forwarder. These MAC entries does not only contain the local MACs of the hosts on the local link but also includes the MAC addresses from the same VLAN on the remote link (i.e., the same virtual link).

All the appointed forwarders will report this sub-TLV messages to the DRB of a LAN link. The information contained in these sub-TLV messages will help the DRB to make more balanced VLAN assignment among the RBridges on the LAN link. Because of host mobility, a former balanced VLAN assignment MAY become unbalanced. If a forwarder's MAC table is running out of use, the DRB can remove some VLANs from it and reassign them to another RBridge as the new forwarder. The number of "MAC Entries of VLANs" SHOULD be constrained by the inter-RBridge link MTU that defaults to 1470 bytes. If the MTU is not big enough to hold all the "MAC Entries of VLANs", the appointed forwarder MAY define its own policy to choose which VLANs it wants the DRB to remove [TBD].

## 6.2. Traffic Bit Rate Report sub-TLV

The appointed forwarder use Traffic Bit Rate Report sub-TLV to report the bandwidth utilization of its port group to the DRB. This sub-TLV has the following format:

```

+-----+
|Type=TrafficRep|                               (1 byte)
+-----+
|  Length      |                               (1 byte)
+-----+
|  DRB Nickname |                               (2 bytes)
+-----+
| Maximum Link Bandwidth |                     (2 bytes)
+-----+

```

Available Link Bandwidth	(2 bytes)
Traffic Bit Rate of VLAN (1)	(4 bytes)
.....	(4 bytes)
Traffic Bit Rate of VLAN (n)	(4 bytes)

where each Load of VLAN is of the form:



```

+---+---+---+---+---+---+---+---+---+---+
| RESV |   VLAN ID   |                      | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+
| Traffic Bit Rate |                      | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+

```

- o Type: Traffic Bit Rate Report sub-TLV.
- o Length: 6+4n bytes, where n is the number of VLANs that the appointed forwarder selects to report their traffic load that egress onto the port group.
- o DRB Nickname: The nickname of the Designated RBridge of the local link.
- o Maximum Link Bandwidth: The maximum bandwidth of the port group attached to the local link.
- o Available Link Bandwidth: The available bandwidth of the port group attached to the local link.
- o RESV: 4 bits that MUST be sent as zero and ignored on receipt.
- o VLAN ID: This field identifies one of the VLANs that assigned to the appointed forwarder.
- o Traffic Bit Rate: The traffic bit rate of the given VLAN onto the local link through the port group of the appointed forwarder.

The appointed forwarder send messages of this sub-TLV to its DRB. The DRB will know the bandwidth utilization of the port group of the appointed forwarder. If the port group of an RBridge attached to the local link is already heavily used, the DRB will refrain from assigning additional VLANs to this RBridge. If an appointed forwarder's port group attached to the local link is congested, its DRB MAY remove some of the VLANs reported in the Traffic Bit Rate Report TLV message and reassign these VLANs to other RBridges attached to the same local link, which will decrease the traffic bit rate via that RBridge. The policy to decide which VLANs to reassign is [TBD].

## 7. Security Considerations

The delivery of the messages types in this document can be protected with the cryptographic mechanism proposed in [RFC5310]. In the future, TRILL MAY define its own secure control message transmission. The new message types introduced in this document can make use of that secure channel.

## 8. IANA Considerations

Two code points of IS-IS sub-TLVs need to be assigned. This work should be done in conjunction with the work of [TRILLtlv].

## 9. References

### 9.1. Normative References

- [TRILLbase] R. Perlman, D. Eastlake, D.G. Dutt, S. Gai and A. Ghanwani, "RBridges: Base Protocol Specification", draft-ietf-trill-rbridge-protocol-16.txt, working in progress.
- [TRILLtlv] D. Eastlake, A. Banerjee, D. Dutt, R. Perlman and A. Ghanwani, "TRILL Use of IS-IS", draft-ietf-trill-adj-02.txt, working in progress.
- [TRILLmib] A. Rijhsinghani, K. Zebrose, "Definitions of Managed Objects for RBridges", draft-ietf-trill-rbridge-mib-02.txt, working in progress.
- [RFC5310] M. Bhatia, V. Manral, T. Li, et al., "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.

### 9.2. Informative References

- [CAMtable] B. Hedlund, "Evolving Data Center Switching", <http://internetworkexpert.s3.amazonaws.com/2010/trill1/TRILL-intro-part1.pdf>
- [SAN] "Configuring an iSCSI Storage Area Network Using Brocade FCX Switches", Brocade CONFIGURATION GUIDE, 2010.
- [VL2] A. Greenberg, J.R. Hamilton, N Jain, et al., "VL2: A scalable and flexible data center network", in Proceedings of ACM SIGCOMM, 2009.
- [TE] M. Roughan, M. Throup, and Y. Zhang, "Traffic Engineering with Estimated Traffic Matrices" , in Proceedings of ACM IMC, 2003.
- [802-1AX] "IEEE Standard for Local and metropolitan area networks - Link Aggregation", IEEE Std 802.1 AX-2008, 3 November 2008.
- [vPC+] "Cisco Nexus 7000 Series NX-OS FabricPath Configuration Guide: FabricPath Interfaces", [http://www.cisco.com/en/US/docs/switches/datacenter/sw/5\\_x/nx-os/fabricpath/](http://www.cisco.com/en/US/docs/switches/datacenter/sw/5_x/nx-os/fabricpath/)

[configuration/guide/fp\\_interfaces.html#wp1666512](#)

[Brocade] Somesh Gupta, Anoop Ghanwani, Phanidhar Koganti, and Shunjia Yu, "Redundant Host Connections in a Routed Network", Patent Application Publication, US 2010/0246388A1, Sep. 30, 2010.

Author's Addresses

Mingui Zhang  
Huawei Technologies Co.,Ltd  
HuaWei Building, No.3 Xinxu Rd., Shang-Di  
Information Industry Base, Hai-Dian District,  
Beijing, 100085 P.R. China

Email: zhangmingui@huawei.com

Dacheng Zhang  
Huawei Technologies Co.,Ltd  
HuaWei Building, No.3 Xinxu Rd., Shang-Di  
Information Industry Base, Hai-Dian District,  
Beijing, 100085 P.R. China

Email: zhangdacheng@huawei.com