

Network Working Group
INTERNET-DRAFT
Intended status: Proposed Standard
Obsoletes: 6326

Donald Eastlake
Huawei
Anoop Ghanwani
Dell
Radia Perlman
Intel
Dinesh Dutt
Ayan Banerjee
Cumulus Networks
July 11, 2012

Expires: January 10, 2013

Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS
<draft-eastlake-isis-rfc6326bis-08.txt>

Abstract

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol provides optimal pair-wise data frame forwarding without configuration in multi-hop networks with arbitrary topology and link technology, and support for multipathing of both unicast and multicast traffic. This document specifies the data formats and code points for the IS-IS extensions to support TRILL. These data formats and code points may also be used by technologies other than TRILL. This document obsoletes RFC 6326.

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 Conventions Used in This Document.....	3
2. TLV and Sub-TLV Extensions to IS-IS for TRILL.....	5
2.1 Group Address TLV.....	5
2.1.1 Group MAC Address Sub-TLV.....	5
2.1.2 Group IPv4 Address Sub-TLV.....	7
2.1.3 Group IPv6 Address Sub-TLV.....	8
2.1.4 Group Labeled MAC Address Sub-TLV.....	8
2.1.5 Group Labeled IPv4 Address Sub-TLV.....	10
2.1.6 Group Labeled IPv6 Address Sub-TLV.....	11
2.2 Multi-Topology-Aware Port Capability Sub-TLVs.....	11
2.2.1 Special VLANs and Flags Sub-TLV.....	12
2.2.2 Enabled-VLANs Sub-TLV.....	13
2.2.3 Appointed Forwarders Sub-TLV.....	14
2.2.4 Port TRILL Version Sub-TLV.....	15
2.2.5 VLANs Appointed Sub-TLV.....	16
2.3 Sub-TLVs for the Router Capability TLV.....	17
2.3.1 TRILL Version Sub-TLV.....	17
2.3.2 Nickname Sub-TLV.....	18
2.3.3 Trees Sub-TLV.....	19
2.3.4 Tree Identifiers Sub-TLV.....	20
2.3.5 Trees Used Identifiers Sub-TLV.....	21
2.3.6 Interested VLANs and Spanning Tree Roots Sub-TLV.....	21
2.3.7 VLAN Group Sub-TLV.....	23
2.3.8 Interested Labels and Spanning Tree Roots Sub-TLV.....	24
2.3.9 RBridge Channel Protocols Sub-TLV.....	26
2.3.10 Affinity Sub-TLV.....	27
2.3.11 Label Group Sub-TLV.....	29
2.4 MTU Sub-TLV of the Extended Reachability TLV.....	30
2.5 TRILL Neighbor TLV.....	30
3. MTU PDUs.....	33
4. Use of Existing PDUs and TLVs.....	34
4.1 TRILL IIH PDUs.....	34
4.2 Area Address.....	34
4.3 Protocols Supported.....	34
4.4 Link State PDUs (LSPs).....	35
4.5 Originating LSP Buffer Size.....	35
5. IANA Considerations.....	36
5.1 TLVs.....	36
5.2 sub-TLVs.....	36
5.3 PDUs.....	37
5.4 Reserved and Capability Bits.....	38
5.5 TRILL Neighbor Record Flags.....	38
6. Security Considerations.....	39
7. Change from RFC 6326.....	40
8. Normative References.....	42
9. Informative References.....	43

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] [RFC6327] provides transparent forwarding in multi-hop networks with arbitrary topology and link technologies using encapsulation with a hop count and link state routing. TRILL provides optimal pair-wise forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. Intermediate Systems (ISs) implementing TRILL are called RBridges (Routing Bridges) or TRILL Switches.

This document, in conjunction with [RFC6165], specifies the data formats and code points for the IS-IS [ISO-10589] [RFC1195] extensions to support TRILL. These data formats and code points may also be used by technologies other than TRILL.

This document obsoletes [RFC6326] which generally corresponded to the base TRILL protocol as it was passed up to the IESG by the TRILL Working Group in 2009. There has been substantial development of TRILL since then. The main changes from [RFC6326] are summarized below and a full list is given in Section 7.

1. Addition of multicast group announcements by IPv4 and IPv6 address.
2. Addition of facilities for announcing capabilities supported.
3. Addition of a tree affinity sub-TLV whereby ISs can request distribution tree association.
4. Addition of control plane support for TRILL Data frame fine-grained labels. This support is independent of the data plane representation.
5. Fix the reported errata [Err2869] in [RFC6326].

Changes herein to TLVs and sub-TLVs specified in [RFC6326] are backwards compatible.

1.1 Conventions Used in This Document

The terminology and acronyms defined in [RFC6325] are used herein with the same meaning.

Additional acronyms and phrases used in this document are:

BVL - Bit Vector Length

BVO - Bit Vector Offset

IIH - IS-IS Hello

IS - Intermediate System. For this document, all relevant intermediate systems are RBridges [RFC6325].

NLPID - Network Layer Protocol Identifier

SNPA - SubNetwork Point of Attachment (MAC Address)

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. TLV and Sub-TLV Extensions to IS-IS for TRILL

This section, in conjunction with [RFC6165], specifies the data formats and code points for the TLVs and sub-TLVs for IS-IS to support the IETF TRILL protocol. Information as to the number of occurrences allowed, such as for a TLV in a PDU or set of PDUs or for a sub-TLV in a TLV, is summarized in Section 5.

2.1 Group Address TLV

The Group Address (GADDR) TLV, IS-IS TLV type 142, is carried in an LSP PDU and carries sub-TLVs that in turn advertise multicast group listeners. The sub-TLVs that advertises listeners are specified below. The sub-TLVs under GADDR constitute a new series of sub-TLV types (see Section 5.2).

GADDR has the following format:

```

+-----+
|Type=GADDR-TLV |                    (1 byte)
+-----+
|  Length      |                    (1 byte)
+-----+
| sub-TLVs...  |
+-----+
```

- o Type: TLV Type, set to GADDR-TLV 142.
- o Length: variable depending on the sub-TLVs carried.
- o sub-TLVs: The Group Address TLV value consists of sub-TLVs formatted as described in [RFC5305].

2.1.1 Group MAC Address Sub-TLV

The Group MAC Address (GMAC-ADDR) sub-TLV is sub-TLV type number 1 within the GADDR TLV. In TRILL, it is used to advertise multicast listeners by MAC address as specified in Section 4.5.5 of [RFC6325]. It has the following format:


```

+---+---+---+---+---+
|Type=GMAC-ADDR |          (1 byte)
+---+---+---+---+---+
|   Length      |          (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  RESV |      Topology-ID      |  (2 bytes)
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  RESV |      VLAN ID          |  (2 bytes)
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|Num Group Recs |          (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     GROUP RECORDS (1)                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     GROUP RECORDS (2)                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     .....                                             |
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     GROUP RECORDS (N)                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

where each group record is of the following form with k=6:

```

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

```

- o Type: GADDR sub-TLV type, set to 1 (GMAC-ADDR).
- o Length: $5 + m + k*n = 5 + m + 6*n$ where m is the number of group records and n is the sum of the number of group and source addresses.
- o RESV: Reserved. 4-bit fields that MUST be sent as zero and ignored on receipt.
- o Topology-ID: This field carries a topology ID [RFC5120] or zero if topologies are not in use.
- o VLAN ID: This carries the 12-bit VLAN identifier for all subsequent MAC addresses in this sub-TLV, or the value zero if no

VLAN is specified.

- o Number of Group Records: A 1-byte unsigned integer that is the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. If this field is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. It then has a 6-byte (48-bit) multicast address followed by 6-byte 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.

The GMAC-ADDR sub-TLV is carried only within a GADDR TLV.

2.1.1.2 Group IPv4 Address Sub-TLV

The Group IPv4 Address (GIP-ADDR) sub-TLV is IS-IS sub-TLV type TBD [2 suggested] within the GADDR TLV. It has the same format as the Group MAC Address sub-TLV described in Section 2.1.1 except that $k=4$. The fields are as follows:

- o Type: sub-TLV Type, set to TBD [2 suggested] (GIP-ADDR).
- o Length: $5 + m + k*n = 5 + m + 4*n$ where m is the number of group records and n is the sum of the number of group and source addresses.
- o Topology-Id: This field carries a topology ID [RFC5120] or zero if topologies are not in use.
- o RESV: Must be sent as zero on transmission and is ignored on receipt.
- o VLAN-ID: This carries a 12-bit VLAN identifier that is valid for all subsequent addresses in this sub-TLV, or the value zero if no VLAN is specified.
- o Number of Group Records: This is of length 1 byte and lists the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. If this field is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. It then has a 4-byte (32-bit) IPv4 Group Address followed by 4-byte source IPv4 addresses. If the number of sources do not fit in a single sub-TLV, it is permitted to have the same group address repeated with different source addresses in another sub-TLV of another instance of the

Group Address TLV.

The GIP-ADDR sub-TLV is carried only within a GADDR TLV.

2.1.3 Group IPv6 Address Sub-TLV

The Group IPv6 Address (GIPV6-ADDR) sub-TLV is IS-IS sub-TLV type TBD [3 suggested] within the GADDR TLV. It has the same format as the Group MAC Address sub-TLV described in Section 2.1.1 except that $k=16$. The fields are as follows:

- o Type: sub-TLV Type, set to TBD [3 suggested] (GIPV6-ADDR).
- o Length: $5 + m + k*n = 5 + m + 16*n$ where m is the number of group records and n is the sum of the number of group and source addresses.
- o Topology-Id: This field carries a topology ID [RFC5120] or zero if topologies are not in use.
- o RESV: Must be sent as zero on transmission and is ignored on receipt.
- o VLAN-ID: This carries a 12-bit VLAN identifier that is valid for all subsequent addresses in this sub-TLV, or the value zero if no VLAN is specified.
- o Number of Group Records: This is of length 1 byte and lists the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. If this field is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. It then has a 16-byte (128-bit) IPv6 Group Address followed by 16-byte source IPv6 addresses. If the number of sources do not fit in a single sub-TLV, it is permitted to have the same group address repeated with different source addresses in another sub-TLV of another instance of the Group Address TLV.

The GIPV6-ADDR sub-TLV is carried only within a GADDR TLV.

2.1.4 Group Labeled MAC Address Sub-TLV

The GMAC-ADDR sub-TLV of the Group Address (GADDR) TLV specified in Section 2.1.1 provides for a VLAN-ID. The Group Labeled MAC Address sub-TLV, below, extends this to a fine-grained label.


```

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

```

where each group record is of the following form with k=6:

```

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

```

- o Type: GADDR sub-TLV Type, set to TBD [4 suggested] (GLMAC-ADDR).
- o Length: $6 + m + k*n = 6 + m + 6*n$ where m is the number of group records and n is the sum of the number of group and source addresses.
- o RESV: Reserved. 4-bit field that MUST be sent as zero and ignored on receipt.
- o Topology-ID: This field carries a topology ID [RFC5120] or zero if topologies are not in use.
- o Label: This carries the 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 unsigned integer that is the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. If this field is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. It then has a 6-byte (48-bit) multicast address followed by 6-byte 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.

The GLMAC-ADDR sub-TLV is carried only within a GADDR TLV.

2.1.5 Group Labeled IPv4 Address Sub-TLV

The Group Labeled IPv4 Address (GLIP-ADDR) sub-TLV is IS-IS sub-TLV type TBD [5 suggested] within the GADDR TLV. It has the same format as the Group Labeled MAC Address sub-TLV described in Section 2.1.4 except that $k=4$. The fields are as follows:

- o Type: sub-TLV Type, set to TBD [5 suggested] (GLIP-ADDR).
- o Length: $6 + m + k*n = 6 + m + 4*n$ where m is the number of group records and n is the sum of the number of group and source addresses.
- o Topology-Id: This field carries a topology ID [RFC5120] or zero if topologies are not in use.
- o RESV: Must be sent as zero on transmission and is ignored on receipt.
- o Label: This carries the fine-grained label identifier for all subsequent IPv4 addresses in this sub-TLV, or the value zero if no label is specified.
- o Number of Group Records: This is of length 1 byte and lists the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. If this field is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. It then has a 4-byte (32-bit) IPv4 Group Address followed by 4-byte source IPv4 addresses. If the number of sources do not fit in a single sub-TLV, it is permitted to have the same group address repeated with different source addresses in another sub-TLV of another instance of the

Group Address TLV.

The GLIP-ADDR sub-TLV is carried only within a GADDR TLV.

2.1.6 Group Labeled IPv6 Address Sub-TLV

The Group Labeled IPv6 Address (GLIPV6-ADDR) sub-TLV is IS-IS sub-TLV type TBD [6 suggested] within the GADDR TLV. It has the same format as the Group Labeled MAC Address sub-TLV described in Section 2.1.4 except that $k=16$. The fields are as follows:

- o Type: sub-TLV Type, set to TBD [6 suggested] (GLIPV6-ADDR).
- o Length: $6 + m + k*n = 6 + m + 16*n$ where m is the number of group records and n is the sum of the number of group and source addresses.
- o Topology-Id: This field carries a topology ID [RFC5120] or zero if topologies are not in use.
- o RESV: Must be sent as zero on transmission and is ignored on receipt.
- o Label: This carries the fine-grained label identifier for all subsequent IPv6 addresses in this sub-TLV, or the value zero if no label is specified.
- o Number of Group Records: This of length 1 byte and lists the number of group records in this sub-TLV.
- o Group Record: Each group record carries the number of sources. If this field is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. It then has a 16-byte (128-bit) IPv6 Group Address followed by 16-byte source IPv6 addresses. If the number of sources do not fit in a single sub-TLV, it is permitted to have the same group address repeated with different source addresses in another sub-TLV of another instance of the Group Address TLV.

The GLIPV6-ADDR sub-TLV is carried only within a GADDR TLV.

2.2 Multi-Topology-Aware Port Capability Sub-TLVs

TRILL makes use of the Multi-Topology-Aware Port Capability (MT-PORT-CAP) TLV as specified in [RFC6165]. The following subsections of

this Section 2.2 specify the sub-TLVs transported by the MT-PORT-CAP TLV for TRILL.

2.2.1 Special VLANs and Flags Sub-TLV

In TRILL, a Special VLANs and Flags (VLAN-Flags) sub-TLV is carried in every IIH PDU. It has the following format:

+---+---+---+---+---+---+---+---+---+---+										
	Type									(1 byte)
+---+---+---+---+---+---+---+---+---+---+										
	Length									(1 byte)
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										
	Port ID									(2 bytes)
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										
	Sender Nickname									(2 bytes)
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										
	AF		AC		VM		BY		Outer.VLAN	(2 bytes)
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										
	TR		R		R		R		Desig.VLAN	(2 bytes)
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										

- o Type: sub-TLV type, set to MT-PORT-CAP VLAN-FLAGS sub-TLV 1.
- o Length: 8.
- o Port ID: An ID for the port on which the enclosing TRILL IIH PDU is being sent as specified in [RFC6325], Section 4.4.2.
- o Sender Nickname: If the sending IS is holding any nicknames as discussed in [RFC6325], Section 3.7, one MUST be included here. Otherwise, the field is set to zero. This field is to support intelligent end stations that determine the egress IS (RBridge) for unicast data through a directory service or the like and that need a nickname for their first hop to insert as the ingress nickname to correctly format a TRILL Data frame (see [RFC6325], Section 4.6.2, point 8). It is also referenced in connection with the VLANs Appointed Sub-TLV (see Section 2.2.5).
- o Outer.VLAN: A copy of the 12-bit outer VLAN ID of the TRILL IIH frame containing this sub-TLV when that frame was sent, as specified in [RFC6325], Section 4.4.5.
- o Desig.VLAN: The 12-bit ID of the Designated VLAN for the link, as specified in [RFC6325], Section 4.2.4.2.

- o AF, AC, VM, BY, and TR: These flag bits have the following meanings when set to one, as specified in the listed section of [RFC6325]:

RFC 6325		
Bit	Section	Meaning if bit is one

AF	4.4.2	Originating IS believes it is appointed forwarder for the VLAN and port on which the containing IIH PDU was sent.
AC	4.9.1	Originating port configured as an access port (TRILL traffic disabled).
VM	4.4.5	VLAN mapping detected on this link.
BY	4.4.2	Bypass pseudonode.
TR	4.9.1	Originating port configured as a trunk port (end-station service disabled).

- o R: Reserved bit. MUST be sent as zero and ignored on receipt.

2.2.2 Enabled-VLANs Sub-TLV

The optional Enabled-VLANs sub-TLV specifies the VLANs enabled at the port of the originating IS on which the containing Hello was sent, as specified in [RFC6325], Section 4.4.2. It has the following format:

```

+---+---+---+---+---+
|      Type      |                               (1 byte)
+---+---+---+---+---+
|      Length     |                               (1 byte)
+---+---+---+---+---+
| RESV | Start VLAN ID | (2 bytes)
+---+---+---+---+---+
| VLAN bit-map....
+---+---+---+---+---+

```

- o Type: sub-TLV type, set to MT-PORT-CAP Enabled-VLANs sub-TLV 2.
- o Length: Variable, minimum 3.
- o RESV: 4 reserved bits that MUST be sent as zero and ignored on receipt.
- o Start VLAN ID: The 12-bit VLAN ID that is represented by the high

order bit of the first byte of the VLAN bit-map.

- o VLAN bit-map: The highest order bit indicates the VLAN equal to the start VLAN ID, the next highest bit indicates the VLAN equal to start VLAN ID + 1, continuing to the end of the VLAN bit-map field.

If this sub-TLV occurs more than once in a Hello, the set of enabled VLANs is the union of the sets of VLANs indicated by each of the Enabled-VLAN sub-TLVs in the Hello.

2.2.3 Appointed Forwarders Sub-TLV

The DRB on a link uses the Appointed Forwarders sub-TLV to inform other ISs on the link that they are the designated VLAN-x forwarder for one or more ranges of VLAN IDs as specified in [RFC6439]. It has the following format:

```

+-----+
|      Type      | (1 byte)
+-----+
|      Length     | (1 byte)
+-----+
| Appointment Information (1) | (6 bytes)
+-----+
| Appointment Information (2) | (6 bytes)
+-----+
| .....          |
+-----+
| Appointment Information (N) | (6 bytes)
+-----+

```

where each appointment is of the form:

```

+-----+
| Appointee Nickname | (2 bytes)
+-----+
| RESV | Start.VLAN | (2 bytes)
+-----+
| RESV | End.VLAN   | (2 bytes)
+-----+

```

- o Type: sub-TLV type, set to MT-PORT-CAP AppointedFwrdrsr sub-TLV 3.
- o Length: 6*n bytes, where there are n appointments.
- o Appointee Nickname: The nickname of the IS being appointed a forwarder.

reserved to indicate support of optional capabilities. A one bit indicates that the flag or capability is supported by the sending IS. Bits in this field MUST be set to zero except as permitted for a capability being advertised or if a hop-by-hop extended header flag is supported.

This sub-TLV, if present, MUST occur in an MT-PORT-CAP TLV in a TRILL IIH. If there is more than one occurrence, the minimum of the supported versions is assumed to be correct and a capability or header flag is assumed to be supported only if indicated by all occurrences. The flags and capabilities for which support can be indicated in this sub-TLV are disjoint from those in the TRILL-VER sub-TLV (Section 2.3.1) so they cannot conflict. The flags and capabilities indicated in this sub-TLV relate to hop-by-hop processing that can differ between the ports of an IS (RBridge), and thus must be advertised in IIHs. For example, a capability requiring cryptographic hardware assist might be supported on some ports and not others. However, the TRILL version is the same as that in the PORT-TRILL-VER sub-TLV and an IS, if it is adjacent to the sending IS of TRILL-VER sub-TLV(s) uses the TRILL version it received in PORT-TRILL-VER sub-TLV(s) in preference to that received in TRILL-VER sub-TLV(s).

2.2.5 VLANs Appointed Sub-TLV

The optional VLANs sub-TLV specifies the VLANs for which a port of the originating IS on which the containing Hello was sent is appointed forwarder. It has the following format:

```

+-----+
|      Type      | (1 byte)
+-----+
|      Length     | (1 byte)
+-----+
| RESV | Start VLAN ID | (2 bytes)
+-----+
| VLAN bit-map....|
+-----+
```

- o Type: sub-TLV type, set to MT-PORT-CAP VLANs-Appointed sub-TLV TBD [8 suggested].
- o Length: Variable, minimum 3.
- o RESV: 4 reserved bits that MUST be sent as zero and ignored on receipt.
- o Start VLAN ID: The 12-bit VLAN ID that is represented by the high

order bit of the first byte of the VLAN bit-map.

- o VLAN bit-map: The highest order bit indicates the VLAN equal to the start VLAN ID, the next highest bit indicates the VLAN equal to start VLAN ID + 1, continuing to the end of the VLAN bit-map field.

If this sub-TLV occurs more than once in a Hello, the originating IS is declaring it believes itself to be appointed forwarder on the port on which the enclosing IIH was sent for the union of the sets of VLANs indicated by each of the VLANs-Appointed sub-TLVs in the Hello.

2.3 Sub-TLVs for the Router Capability TLV

The Router Capability TLV is specified in [RFC4971]. All of the sub-sections of this Section 2.3 below specify sub-TLVs that can be carried in the Router Capability TLV for TRILL which in turn is carried only by LSPs.

2.3.1 TRILL Version Sub-TLV

The TRILL Version (TRILL-VER) sub-TLV indicates the maximum version of the TRILL standard supported and the support of optional capabilities by the originating IS. By implication, lower versions are also supported. If this sub-TLV is missing, it is assumed that the originating IS only supports the base version (version zero) of the protocol [RFC6325] and no optional capabilities indicated by this sub-TLV are supported.

```

+-----+
| Type                | (1 byte)
+-----+
| Length              | (1 byte)
+-----+
| Max-version         | (1 byte)
+-----+-----+
| Capabilities and Header Flags Supported | (4 bytes)
+-----+-----+
0          1          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 0 1

```

- o Type: Router Capability sub-TLV type, set to 13 (TRILL-VER).
- o Length: 5.

- o Max-version: A one byte unsigned integer set to maximum version supported.
- o Capabilities and Header Flags Supported: A bit vector of 32 bits numbered 0 through 31 in network order. Bits 14 through 31 indicate that the corresponding TRILL Header extended flags [ExtendHeader] are supported. Bits 0 through 13 are reserved to indicate support of optional capabilities. A one bit indicates that the originating IS supports the flag or capability. For example, support of multi-level TRILL IS-IS [MultiLevel]. Bits in this field MUST be set to zero except as permitted for a capability being advertised or an extended header flag supported.

This sub-TLV, if present, MUST occur in a Router Capabilities TLV in the LSP number zero for the originating IS. If found in other fragments, it is ignored. If there is more than one occurrence in LSP number zero, the minimum of the supported versions is assumed to be correct and an extended header flag or capability is assumed to be supported only if indicated by all occurrences. The flags and capabilities supported bits in this sub-TLV are disjoint from those in the PORT-TRILL-VER sub-TLV (Section 2.2.4) so they cannot conflict. However, the TRILL version is the same as that in the PORT-TRILL-VER sub-TLV and an IS that is adjacent to the originating IS of TRILL-VER sub-TLV(s) uses the TRILL version it received in PORT-TRILL-VER sub-TLV(s) in preference to that received in TRILL-VER sub-TLV(s).

2.3.2 Nickname Sub-TLV

The Nickname (NICKNAME) Router Capability sub-TLV carries information about the nicknames of the originating IS, along with information about its priority to hold those nicknames as specified in [RFC6325], Section 3.7.3. Multiple instances of this sub-TLV may be carried.

```

+---+---+---+---+---+
|Type = NICKNAME|                                     (1 byte)
+---+---+---+---+---+
|   Length      |                                     (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     NICKNAME RECORDS (1)                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     NICKNAME RECORDS (2)                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     .....                                                 |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     NICKNAME RECORDS (N)                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```


where each nickname record is of the form:

```

+---+---+---+---+---+
| Nickname.Pri |                               (1 byte)
+---+---+---+---+---+
|   Tree Root Priority   | (2 byte)
+---+---+---+---+---+
|           Nickname     | (2 bytes)
+---+---+---+---+---+

```

- o Type: Router Capability sub-TLV type, set to 6 (NICKNAME).
- o Length: 5*n, where n is the number of nickname records present.
- o Nickname.Pri: An 8-bit unsigned integer priority to hold a nickname as specified in Section 3.7.3 of [RFC6325].
- o Tree Root Priority: This is an unsigned 16-bit integer priority to be a tree root as specified in Section 4.5 of [RFC6325].
- o Nickname: This is an unsigned 16-bit integer as specified in Section 3.7 of [RFC6325].

2.3.3 Trees Sub-TLV

Each IS providing TRILL service uses the TREES sub-TLV to announce three numbers related to the computation of distribution trees as specified in Section 4.5 of [RFC6325]. Its format is as follows:

```

+---+---+---+---+---+
| Type = TREES |                               (1 byte)
+---+---+---+---+---+
| Length |                               (1 byte)
+---+---+---+---+---+
| Number of trees to compute | (2 byte)
+---+---+---+---+---+
| Maximum trees able to compute | (2 byte)
+---+---+---+---+---+
| Number of trees to use | (2 byte)
+---+---+---+---+---+

```

- o Type: Router Capability sub-TLV type, set to 7 (TREES).
- o Length: 6.
- o Number of trees to compute: An unsigned 16-bit integer as specified in Section 4.5 of [RFC6325].

- o Maximum trees able to compute: An unsigned 16-bit integer as specified in Section 4.5 of [RFC6325].
- o Number of trees to use: An unsigned 16-bit integer as specified in Section 4.5 of [RFC6325].

2.3.4 Tree Identifiers Sub-TLV

The tree identifiers (TREE-RT-IDs) sub-TLV is an ordered list of nicknames. When originated by the IS that has the highest priority to be a tree root, it lists the distribution trees that the other ISs are required to compute as specified in Section 4.5 of [RFC6325]. If this information is spread across multiple sub-TLVs, the starting tree number is used to allow the ordered lists to be correctly concatenated. The sub-TLV format is as follows:

```

+-----+
|Type=TREE-RT-IDs|          (1 byte)
+-----+
|  Length  |                (1 byte)
+-----+
|Starting Tree Number|      (2 bytes)
+-----+
|  Nickname (K-th root)  |   (2 bytes)
+-----+
|  Nickname (K+1 - th root) | (2 bytes)
+-----+
|  Nickname (...)  |
+-----+

```

- o Type: Router Capability sub-TLV type, set to 8 (TREE-RT-IDs).
- o Length: $2 + 2*n$, where n is the number of nicknames listed.
- o Starting Tree Number: This identifies the starting tree number of the nicknames that are trees for the domain. This is set to 1 for the sub-TLV containing the first list. Other Tree-Identifiers sub-TLVs will have the number of the starting list they contain. In the event the same tree identifier can be computed from two such sub-TLVs and they are different, then it is assumed that this is a transient condition that will get cleared. During this transient time, such a tree SHOULD NOT be computed unless such computation is indicated by all relevant sub-TLVs present.
- o Nickname: The nickname at which a distribution tree is rooted.

2.3.5 Trees Used Identifiers Sub-TLV

This Router Capability sub-TLV has the same structure as the Tree Identifiers sub-TLV specified in Section 2.3.4. The only difference is that its sub-TLV type is set to 9 (TREE-USE-IDs), and the trees listed are those that the originating IS wishes to use as specified in [RFC6325], Section 4.5.

2.3.6 Interested VLANs and Spanning Tree Roots Sub-TLV

The value of this Router Capability sub-TLV consists of a VLAN range and information in common to all of the VLANs in the range for the originating IS. This information consists of flags, a variable length list of spanning tree root bridge IDs, and an appointed forwarder status lost counter, all as specified in the sections of [RFC6325] listed with the respective information items below.

In the set of LSPs originated by an IS, the union of the VLAN ranges in all occurrences of this sub-TLV MUST be the set of VLANs for which the originating IS is appointed forwarder on at least one port, and the VLAN ranges in multiple VLANs sub-TLVs for an IS MUST NOT overlap unless the information provided about a VLAN is the same in every instance. However, as a transient state these conditions may be violated. If a VLAN is not listed in any INT-VLAN sub-TLV for an IS, that IS is assumed to be uninterested in receiving traffic for that VLAN. If a VLAN appears in more than one INT-VLAN sub-TLV for an IS with different information in the different instances, the following apply:

- If those sub-TLVs provide different nicknames, it is unspecified which nickname takes precedence.
- The largest appointed forwarder status lost counter, using serial number arithmetic [RFC1982], is used.
- The originating IS is assumed to be attached to a multicast IPv4 router for that VLAN if any of the INT-VLAN sub-TLVs assert that it is so connected and similarly for IPv6 multicast router attachment.
- The root bridge lists from all of the instances of the VLAN for the originating IS are merged.

To minimize such occurrences, wherever possible, an implementation SHOULD advertise the update to an interested VLAN and Spanning Tree Roots sub-TLV in the same LSP fragment as the advertisement that it replaces. Where this is not possible, the two affected LSP fragments should be flooded as an atomic action. An IS that receives an update to an existing interested VLAN and Spanning Tree Roots sub-TLV can minimize the potential disruption associated with the update by employing a hold-down timer prior to processing the update so as to

allow for the receipt of multiple LSP fragments associated with the same update prior to beginning processing.

The sub-TLV layout is as follows:

```

+-----+
|Type = INT-VLAN|                               (1 byte)
+-----+
|  Length      |                               (1 byte)
+-----+
|  Nickname     |                               (2 bytes)
+-----+
| Interested VLANs |                               (4 bytes)
+-----+
| Appointed Forwarder Status Lost Counter | (4 bytes)
+-----+
|      Root Bridges      | (6*n bytes)
+-----+

```

- o Type: Router Capability sub-TLV type, set to 10 (INT-VLAN).
- o Length: 10 + 6*n, where n is the number of root bridge IDs.
- o Nickname: As specified in [RFC6325], Section 4.2.4.4, this field may be used to associate a nickname held by the originating IS with the VLAN range indicated. When not used in this way, it is set to zero.
- o Interested VLANs: The Interested VLANs field is formatted as shown below.

0	1	2	3	4 - 15	16 - 19	20 - 31
M4	M6	R	R	VLAN.start	RESV	VLAN.end

- 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 VLAN in the indicated range as specified in [RFC6325], Section 4.2.4.4, item 5.1.
- R, RESV: These reserved bits MUST be sent as zero and are ignored on receipt.
- VLAN.start and VLAN.end: This VLAN ID range is inclusive. Setting both VLAN.start and VLAN.end to the same value indicates a range of one VLAN ID. If VLAN.start is not equal to VLAN.end and VLAN.start is 0x000, the sub-TLV is interpreted as if VLAN.start was 0x001. If VLAN.start is not equal to VLAN.end

and VLAN.end is 0xFFFF, the sub-TLV is interpreted as if VLAN.end was 0xFFFE. If VLAN.end is less than VLAN.start, the sub-TLV is ignored. If both VLAN.start and VLAN.end are 0x000 or both are 0xFFFF, the sub-TLV is ignored.

- o Appointed Forwarder Status Lost Counter: This is a count of how many times a port that was appointed forwarder for the VLANs in the range given has lost the status of being an appointed forwarder for some port as discussed in Section 4.8.3 of [RFC6325]. It is initialized to zero at an IS when the zeroth LSP sequence number is initialized. No special action need be taken at rollover; the counter just wraps around.
- o Root Bridges: The list of zero or more spanning tree root bridge IDs is the set of root bridge IDs seen for all ports for which the IS is appointed forwarder for the VLANs in the specified range as discussed in [RFC6325], Section 4.9.3.2. While, of course, at most one spanning tree root could be seen on any particular port, there may be multiple ports in the same VLANs connected to different bridged LANs with different spanning tree roots.

An INT-VLAN sub-TLV asserts that the information provided (multicast router attachment, appointed forwarder status lost counter, and root bridges) is the same for all VLANs in the range specified. If this is not the case, the range MUST be split into subranges meeting this criteria. It is always safe to use sub-TLVs with a "range" of one VLAN ID, but this may be too verbose.

2.3.7 VLAN Group Sub-TLV

The VLAN Group Router Capability sub-TLV consists of two or more VLAN IDs as specified in [RFC6325], Section 4.8.4. This sub-TLV indicates that shared VLAN learning is occurring at the originating IS between the listed VLANs. It is structured as follows:

```

+---+---+---+---+---+
|Type=VLAN-GROUP|          (1 byte)
+---+---+---+---+---+
|  Length      |          (1 byte)
+---+---+---+---+---+
| RESV  | Primary VLAN ID  | (2 bytes)
+---+---+---+---+---+
| RESV  | Secondary VLAN ID | (2 bytes)
+---+---+---+---+---+
| more Secondary VLAN IDs ... (2 bytes each)
+---+---+---+---+---+

```

- o Type: Router Capability sub-TLV type, set to 14 (VLAN-GROUP).

- o Length: $4 + 2*n$, where n is the number of secondary VLAN ID fields beyond the first. n MAY be zero.
- o RESV: a 4-bit field that MUST be sent as zero and ignored on receipt.
- o Primary VLAN ID: This identifies the primary VLAN ID.
- o Secondary VLAN ID: This identifies a secondary VLAN in the VLAN Group.
- o more Secondary VLAN IDs: zero or more byte pairs, each with the top 4 bits as a RESV field and the low 12 bits as a VLAN ID.

2.3.8 Interested Labels and Spanning Tree Roots Sub-TLV

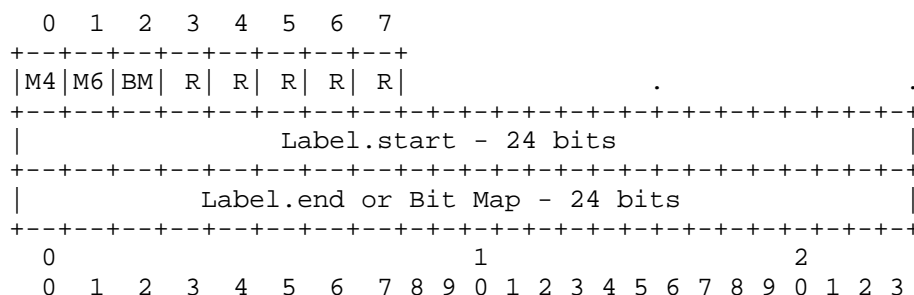
An IS that can handle fine-grained labeling announces its fine-grained label connectivity and related information in the "Interested Labels and Bridge Spanning Tree Roots sub-TLV" (INT-LABEL) which is a variation of the "Interested VLANs and Spanning Tree Roots sub-TLV" (INT-VLAN) structured as below.

```

+-----+-----+
|Type= INT-LABEL|                               (1 byte)
+-----+-----+
|   Length      |                               (1 byte)
+-----+-----+-----+-----+-----+-----+
|   Nickname     |                               (2 bytes)
+-----+-----+-----+-----+-----+-----+...+-----+
|   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 [15 suggested] (INT-LABEL).
- o Length: $11 + 6*n$ where n is the number of root bridge IDs.
- o Nickname: This field may be used to associate a nickname held by the originating IS with the Labels indicated. When not used in this way, it is set to zero.
- o Interested Labels: The Interested Labels field is seven bytes long and formatted as shown below.



- M4, M6: These bits indicate, respectively, that there is an IPv4 or IPv6 multicast router on a link to which the originating IS is appointed forwarder for the VLAN corresponding to every label in the indicated range.
 - BM: If the BM (Bit Map) bit is zero, the last three bytes of the Interested Labels is a Label.end label number. If the BM bit is one, those bytes are a bit map as described below.
 - R: These reserved bits MUST be sent as zero and are ignored on receipt.
 - Label.start and Label.end: If the BM bit is zero: This fine-grained label ID range is inclusive. These fields are treated as unsigned integers. Setting them both to that same label ID value indicates a range of one label ID. If Label.end is less than Label.start, the sub-TLV is ignored.
 - Label.start and Bit Map: If the BM bit is one: The fine-grained labels that the IS is interested in are indicated by a 24-bit bit map. The interested labels are the Label.start number plus the bit number of each one bit in the bit map. So, if bit zero of the bit map is a one, the IS is interested in the label with value Label.start and if bit 23 of the bit map is a one, the IS is interested in the label with value Label.start+23.
- o Appointed Forwarder Status Lost Counter: This is a count of how many times a port that was appointed forwarder for a VLAN mapping to the fine-grained label in the range or bit map given has lost the status of being an appointed forwarder as discussed in Section 4.8.3 of [RFC6325]. It is initialized to zero at an IS when the zeroth LSP sequence number is initialized. No special action need be taken at rollover; the counter just wraps around.
 - o Root Bridges: The list of zero or more spanning tree root bridge IDs is the set of root bridge IDs seen for all ports for which the IS is appointed forwarder for a VLAN mapping to the fine-grained label in the specified range or bit map. (See [RFC6325], Section 4.9.3.2.) While, of course, at most one spanning tree root could

be seen on any particular port, there may be multiple relevant ports connected to different bridged LANs with different spanning tree roots.

An INT-LABEL sub-TLV asserts that the information provided (multicast router attachment, appointed forwarder status lost counter, and root bridges) is the same for all labels specified. If this is not the case, the sub-TLV MUST be split into subranges and/or separate bit maps meeting this criteria. It is always safe to use sub-TLVs with a "range" of one VLAN ID, but this may be too verbose.

2.3.9 RBridge Channel Protocols Sub-TLV

An IS announces the RBridge Channel protocols [Channel] it supports through use of this sub-TLV.

```

+---+---+---+---+---+
|Type=RBCHANNELS|                               (1 byte)
+---+---+---+---+---+
|   Length   |                               (1 byte)
+---+---+---+---+---+---+---+---+---+---+---+...
|   Zero or more bit vectors   |               (variable)
+---+---+---+---+---+

```

- o Type: RBridge Channel Protocols, set to TBD [16 suggested] (RBCHANNELS).
- o Length: variable.
- o Bit Vectors: Zero or more byte-aligned bit vectors where a one bit indicates support of a particular RBridge Channel protocol. Each byte-aligned bit vector is formatted as follows:

```

| 0  1  2  3  4  5  6  7| 8  9 10 11 12 13 14 15|
+---+---+---+---+---+---+---+---+---+---+---+---+
| Bit Vector Length |      Bit Vector Offset      |
+---+---+---+---+---+---+---+---+---+---+---+---+
|   bits   |
+---+---+---+---+---+

```

The bit vector length (BVL) is a seven bit unsigned integer field giving the number of bytes of bit vector. The bit vector offset (BVO) is a nine bit unsigned integer field.

The bits in each bit vector are numbered in network order, the high order bit of the first byte of bits being bit 0 + 8*BVO, the low order bit of that byte being 7 + 8*BVO, the high order bit of the second byte being 8 + 8*BVO, and so on for BVL bytes. An RBridge

Channel protocols-supported bit vector MUST NOT extend beyond the end of the value in the sub-TLV in which it occurs. If it does, it is ignored. If multiple byte-aligned bit vectors are present in one such sub-TLV, their representations are contiguous, the BVL field for the next starting immediately after the last byte of bits for the previous bit vector. The one or more bit vectors present MUST exactly fill the sub-TLV value. If there are one or two bytes of value left over, they are ignored; if more than two, an attempt is made to parse them as one or more bit vectors.

If different bit vectors overlap in the protocol number space they refer to and they have inconsistent bit values for a channel protocol, support for the protocol is assumed if any of these bit vectors has a 1 for that protocol.

The absence of any occurrences of this sub-TLV in the LSP for an IS implies that that IS does not support the RBridge Channel facility.

To avoid wasted space, trailing bit vector zero bytes SHOULD be eliminated by reducing BVL, any null bit vectors (ones with BVL equal to zero) eliminated, and generally the most compact encoding used. For example, support for channel protocols 1 and 32 could be encoded as

```
BVL = 5
BVO = 0
0b01000000
0b00000000
0b00000000
0b00000000
0b10000000
```

or as

```
BVL = 1
BVO = 0
0b01000000
BLV = 1
BVO = 4
0b10000000
```

The first takes 7 bytes while the second takes only 6 and thus the second would be preferred.

2.3.10 Affinity Sub-TLV

Association of an IS to a multi-destination distribution tree through a specific path is accomplished by using the tree Affinity sub-TLV.

The announcement of an Affinity sub-TLV by RB1 with the nickname of RB2 as the first part of an Affinity Record in the sub-TLV value is a request by RB1 that all ISes in the campus connect RB2 as a child of RB1 when calculating any of the trees listed in that Affinity Record.

AFFINITY is a sub-TLV of Router capability TLV (#242) [RFC4971] with the structure shown below.

```

+---+---+---+---+---+
| Type=AFFINITY |           (1 byte)
+---+---+---+---+---+
| Length         |           (1 byte)
+---+---+---+---+---+
|                AFFINITY RECORD 1                |
+---+---+---+---+---+
|                AFFINITY RECORD 2                |
+---+---+---+---+---+
|                .....                            |
+---+---+---+---+---+
|                AFFINITY RECORD N                |
+---+---+---+---+---+

```

Where each AFFINITY RECORD is structured as follows:

```

+---+---+---+---+---+---+---+---+
| Nickname        | (2 bytes)
+---+---+---+---+---+---+---+---+
| Affinity Flags  | (1 byte)
+---+---+---+---+---+
| Number of trees | (1 byte)
+---+---+---+---+---+---+---+---+
| Tree-num of 1st root | (2 bytes)
+---+---+---+---+---+---+---+---+
| Tree-num of 2nd root | (2 bytes)
+---+---+---+---+---+---+---+---+
|                .....                            |
+---+---+---+---+---+---+---+---+
| Tree-num of Nth root | (2 bytes)
+---+---+---+---+---+---+---+---+

```

- o Type: Router Capability sub-TLV type, set to TBD (AFFINITY).
- o Length: 1 + size of all Affinity Records included, where an Affinity Record listing n tree roots is 3+2*n bytes long.
- o Nickname: 16-bit nickname of the IS whose associations to the multi-destination trees listed in the Affinity Record are through the originating IS.
- o Affinity Flags: 8 bits reserved for future needs to provide

additional information about the affinity being announced. MUST be sent as zero and ignored on receipt.

- o Number of trees: A one byte unsigned integer giving the number of trees for which affinity is being announced by this Affinity Record.
- o Tree-num of roots: The tree numbers of the distribution trees this Affinity Record is announcing.

There is no need for a field giving the number of Affinity Records as this can be determined by processing those records.

2.3.11 Label Group Sub-TLV

The Label Group Router Capability sub-TLV consists of two or more Label IDs. This sub-TLV indicates that shared Label MAC address learning is occurring at the announcing IS between the listed Labels. It is structured as follows:

```

+-----+
|Typ=LABEL-GROUP|                               (1 byte)
+-----+
|  Length      |                               (1 byte)
+-----+
| Primary Label ID |                           (3 bytes)
+-----+
| Secondary Label ID |                         (3 bytes)
+-----+
| more Secondary Label IDs ...                 (3 bytes each)
+-----+

```

- o Type: Router Capability sub-TLV type, set to TBD (LABEL-GROUP).
- o Length: $6 + 3*n$, where n is the number of secondary VLAN ID fields beyond the first. n MAY be zero.
- o Primary Label ID: This identifies the primary Label ID.
- o Secondary Label ID: This identifies a secondary Label in the Label Group.
- o more Secondary Label IDs: zero or more byte triples, each with a Label ID.

2.4 MTU Sub-TLV of the Extended Reachability TLV

The MTU sub-TLV is used to optionally announce the MTU of a link as specified in [RFC6325], Section 4.2.4.4. It occurs within the Extended Reachability TLV (type 22).

```

+-----+
| Type = MTU | (1 byte)
+-----+
| Length | (1 byte)
+-----+
| F | RESV | (1 byte)
+-----+
| MTU | (2 bytes)
+-----+

```

- o Type: Extended Reachability sub-TLV type, set to MTU sub-TLV 28.
- o Length: 3.
- o F: Failed. This bit is a one if MTU testing failed on this link at the required campus-wide MTU.
- o RESV: 7 bits that MUST be sent as zero and ignored on receipt.
- o MTU: This field is set to the largest successfully tested MTU size for this link, or zero if it has not been tested, as specified in Section 4.3.2 of [RFC6325].

2.5 TRILL Neighbor TLV

The TRILL Neighbor TLV is used in TRILL IIH PDUs (see Section 4.1 below) in place of the IS Neighbor TLV, as specified in Section 4.4.2.1 of [RFC6325] and in [RFC6327]. The structure of the TRILL Neighbor TLV is as follows:


```

+---+---+---+---+---+
|      Type      |                               (1 byte)
+---+---+---+---+---+
|      Length     |                               (1 byte)
+---+---+---+---+---+
|S|L|R|  SIZE   |                               (1 byte)
+---+---+---+---+---+
|                               Neighbor RECORDS (1)                               |
+---+---+---+---+---+
|                               Neighbor RECORDS (2)                               |
+---+---+---+---+---+
|                               .....                               |
+---+---+---+---+---+
|                               Neighbor RECORDS (N)                               |
+---+---+---+---+---+

```

The information present for each neighbor is as follows:

```

+---+---+---+---+---+
|F|O|  RESV      |                               (1 bytes)
+---+---+---+---+---+
|      MTU       |                               (2 bytes)
+---+---+---+---+---+
|      SNPA (MAC Address)                               | (SIZE bytes)
+---+---+---+---+---+

```

- o Type: TLV Type, set to TRILL Neighbor TLV 145.
- o Length: $1 + (SIZE+3)*n$, where n is the number of neighbor records, which may be zero.
- o S: Smallest flag. If this bit is a one, then the list of neighbors includes the neighbor with the smallest MAC address considered as an unsigned integer.
- o L: Largest flag. If this bit is a one, then the list of neighbors includes the neighbor with the largest MAC address considered as an unsigned integer.
- o R, RESV: These bits are reserved and MUST be sent as zero and ignored on receipt.
- o SIZE: The SNPA size as an unsigned integer in bytes except that 6 is encoded as zero. An actual size of zero is meaningless and cannot be encoded. The meaning of the value 6 in this field is reserved and TRILL Neighbor TLVs received with a SIZE of 6 are ignored. The SIZE is inherent to the technology of a link and is fixed for all TRILL Neighbor TLVs on that link but may vary between different links in the campus if those links are different technologies. For example, 6 for EUI-48 SNPAs or 8 for EUI-64

SNPAs [RFC5342]. (The SNPA size on the various links in a TRILL campus is independent of the System ID size.)

- o F: failed. This bit is a one if MTU testing to this neighbor failed at the required campus-wide MTU (see [RFC6325], Section 4.3.1).
- o O: OOMF. This bit is a one if the IS sending the enclosing TRILL Neighbor TLV is willing to offer the Overload Originated Multi-destination Frame (OOMF) service [ClearCorrect] to the IS whose port has the SNPA in the enclosing Neighbor RECORD.
- o MTU: This field is set to the largest successfully tested MTU size for this neighbor or to zero if it has not been tested.
- o SNPA: Sub-Network Point of Attachment (MAC address) of the neighbor.

As specified in [RFC6327] and Section 4.4.2.1 of [RFC6325], all MAC addresses may fit into one TLV, in which case both the S and L flags would be set to one in that TLV. If the MAC addresses don't fit into one TLV, the highest MAC address in a TRILL Neighbor TLV with the L flag zero MUST also appear as a MAC address in some other TRILL Neighbor TLV (possibly in a different TRILL IIH PDU). Also, the lowest MAC address in a TRILL Neighbor TLV with the S flag zero MUST also appear in some other TRILL Neighbor TLV (possibly in a different TRILL IIH PDU). If an IS believes it has no neighbors, it MUST send a TRILL Neighbor TLV with an empty list of neighbor RECORDS, which will have both the S and L bits on.

3. MTU PDUs

The IS-IS MTU-probe and MTU-ack PDUs are used to optionally determine the MTU on a link between ISs as specified in Section 4.3.2 of [RFC6325] and in [RFC6327].

The MTU PDUs have the IS-IS PDU common header (up through the Maximum Area Addresses byte) with PDU Type numbers as indicated in Section 5. They also have a common fixed MTU PDU header as shown below that is $8 + 2 \times (\text{ID Length})$ bytes long, 20 bytes in the case of the usual 6-bytes System IDs.

```

+---+---+---+---+---+---+---+---+---+---+
|   PDU Length                               | (2 bytes)
+---+---+---+---+---+---+---+---+---+---+.....+---+
|   Probe ID                               (6 bytes)   |
+---+---+---+---+---+---+---+---+---+---+.....+---+
|   Probe Source ID                         (ID Length bytes) |
+---+---+---+---+---+---+---+---+---+---+.....+---+
|   Ack Source ID                         (ID Length bytes) |
+---+---+---+---+---+---+---+---+---+---+.....+---+

```

As with other IS-IS PDUs, the PDU length gives the length of the entire IS-IS packet starting with and including the IS-IS common header.

The Probe ID field is an opaque 48-bit quantity set by the IS issuing an MTU-probe and copied by the responding IS into the corresponding MTU-ack. For example, an IS creating an MTU-probe could compose this quantity from a port identifier and probe sequence number relative to that port.

The Probe Source ID is set by an IS issuing an MTU-probe to its System ID and copied by the responding IS into the corresponding MTU-ack. The Ack Source ID is set to zero in MTU-probe PDUs and ignored on receipt. An IS issuing an MTU-ack sets the Ack Source ID field to its System ID. The System ID length is usually 6 bytes but could be a different value as indicated by the ID Length field in the IS-IS PDU Header.

The TLV area follows the MTU PDU header area. This area MAY contain an Authentication TLV and MUST be padded to the exact size being tested with the Padding TLV. Since the minimum size of the Padding TLV is 2 bytes, it would be impossible to pad to exact size if the total length of the required information bearing fixed fields and TLVs added up to 1 byte less than the desired length. However, the length of the fixed fields and substantive TLVs for MTU PDUs is expected to be quite small compared with their minimum length (minimum 1470-byte MTU on an 802.3 link, for example), so this should not be a problem.

4. Use of Existing PDUs and TLVs

The sub-sections below provide details of TRILL use of existing PDUs and TLVs.

4.1 TRILL IIH PDUs

The TRILL IIH PDU is the variation of the LAN IIH PDU used by the TRILL protocol. Section 4.4 of the TRILL standard [RFC6325] and [RFC6327] specify the contents of the TRILL IIH and how its use in TRILL differs from Layer 3 LAN IIH PDU use. The adjacency state machinery for TRILL neighbors is specified in detail in [RFC6327].

In a TRILL IIH PDU, the IS-IS common header and the fixed PDU Header are the same as a Level 1 LAN IIH PDU. The Maximum Area Addresses octet in the common header MUST be set to 0x01.

The IS-IS Neighbor TLV (6) is not used in a TRILL IIH and is ignored if it appears there. Instead, TRILL IIH PDUs use the TRILL Neighbor TLV (see Section 2.5).

4.2 Area Address

TRILL uses a fixed zero Area Address as specified in [RFC6325], Section 4.2.3. This is encoded in a 4-byte Area Address TLV (TLV #1) as follows:

```

+---+---+---+---+---+---+---+---+---+
| 0x01, Area Address Type | (1 byte)
+---+---+---+---+---+---+---+---+---+
| 0x02, Length of Value | (1 byte)
+---+---+---+---+---+---+---+---+---+
| 0x01, Length of Address | (1 byte)
+---+---+---+---+---+---+---+---+---+
| 0x00, zero Area Address | (1 byte)
+---+---+---+---+---+---+---+---+---+

```

4.3 Protocols Supported

NLPID (Network Layer Protocol ID) 0xC0 has been assigned to TRILL [RFC6328]. A Protocols Supported TLV (#129, [RFC1195]) including that value MUST appear in TRILL IIH PDUs and LSP number zero PDUs.

4.4 Link State PDUs (LSPs)

A number zero LSP MUST NOT be originated larger than 1470 bytes but a larger number zero LSP successfully received MUST be processed and forwarded normally.

4.5 Originating LSP Buffer Size

The originatingLSPBufferSize TLV (#14) MUST be in LSP number zero; however, if found in other LSP fragments, it is processed normally. Should there be more than one originatingLSPBufferSize TLV for an IS, the minimum size, but not less than 1470, is used.

5. IANA Considerations

This section give IANA Considerations for the TLVs, sub-TLVs, and PDUs specified herein.

5.1 TLVs

This document specifies two IS-IS TLV types -- namely, the Group Address TLV (GADDR-TLV, type 142) and the TRILL Neighbor TLV (type 145). The PDUs in which these TLVs are permitted for TRILL are shown in the table below along with the section of this document where they are discussed. The final "NUMBER" column indicates the permitted number of occurrences of the TLV in their PDU, or set of PDUs in the case of LSP, which in these two cases is "*" indicating that the TLV MAY occur 0, 1, or more times.

IANA has registered these two code points in the IANA IS-IS TLV registry (ignoring the "Section" and "NUMBER" columns, which are irrelevant to that registry).

	Section	TLV	IIH	LSP	SNP	Purge	NUMBER
	=====	===	===	===	===	=====	=====
GADDR-TLV	2.1	142	-	X	-	-	*
TRILL Neighbor TLV	2.5	145	X	-	-	-	*

5.2 sub-TLVs

This document specifies a number of sub-TLVs including 12 new sub-TLVs. The TLVs in which these sub-TLVs occur are shown in the second table below along with the section of this document where they are discussed. The TLVs within which these sub-TLVs can occur are determined by the presence of an "X" in the relevant column as shown in the first table below.

Column Head	TLV	RFCref	TLV Name
=====	=====	=====	=====
Grp. Adr.	142	This doc	Group Address
MT Port	143	6165	MT-PORT-CAP
Rtr. Cap	242	4971	Router CAPABILITY
Ext. Reach	22	5305	Extended IS Reachability

The final "NUMBER" column below indicates the permitted number of occurrences of the sub-TLV cumulatively within all occurrences of their TLV in that TLV's carrying PDU (or set of PDUs in the case of LSP), as follows:

0-1 = MAY occur zero or one times.

1 = MUST occur exactly once. If absent, the PDU is ignored. If it occurs more than once, results are unspecified.

* = MAY occur 0, 1, or more times.

The values in the "Section" and "NUMBER" columns are irrelevant to the IANA sub-registries.

Name	Section	sub- TLV#	Grp. Adr.	MT Port	Rtr. Cap.	Ext. Reach	NUMBER
=====							
GMAC-ADDR	2.1.1	1	X	-	-	-	*
GIP-ADDR	2.1.2	TBD[2]	X	-	-	-	*
GIPV6-ADDR	2.1.3	TBD[3]	X	-	-	-	*
GLMAC-ADDR	2.1.4	TBD[4]	X	-	-	-	*
GLIP-ADDR	2.1.5	TBD[5]	X	-	-	-	*
GLIPV6-ADDR	2.1.6	TBD[6]	X	-	-	-	*
VLAN-FLAGS	2.2.1	1	-	X	-	-	1
Enabled-VLANs	2.2.2	2	-	X	-	-	*
AppointedFwrdrs	2.2.3	3	-	X	-	-	*
PORT-TRILL-VER	2.2.4	TBD[7]	-	X	-	-	0-1
VLANs-Appointed	2.2.5	TBD[8]	-	X	-	-	*
NICKNAME	2.3.2	6	-	-	X	-	*
TREES	2.3.3	7	-	-	X	-	0-1
TREE-RT-IDs	2.3.4	8	-	-	X	-	*
TREE-USE-IDs	2.3.5	9	-	-	X	-	*
INT-VLAN	2.3.6	10	-	-	X	-	*
TRILL-VER	2.3.1	13	-	-	X	-	0-1
VLAN-GROUP	2.3.7	14	-	-	X	-	*
INT-LABEL	2.3.8	TBD[15]	-	-	X	-	*
RBCHANNELS	2.3.9	TBD[16]	-	-	X	-	*
AFFINITY	2.3.10	TBD[17]	-	-	X	-	*
LABEL-GROUP	2.3.11	TBD[18]	-	-	X	-	*
MTU	2.4	28	-	-	-	X	0-1
=====							
Name	Section	sub- TLV#	Grp. Adr.	MT Port	Rtr. Cap.	Ext. Reach	NUMBER

5.3 PDUs

The IS-IS PDUs registry remains as established in [RFC6326] except that the references to [RFC6326] are updated to reference this document.

5.4 Reserved and Capability Bits

Any reserved bits (R) or bits in reserved fields (RESV) or the capabilities bits in the PORT-TRILL-VER and TRILL-VER sub-TLVs, which are specified herein as "MUST be sent as zero and ignored on receipt" or the like, are allocated based on IETF Review [RFC5226].

Two sub-registries are created within the TRILL Parameters Registry as follows:

Sub-Registry Name: TRILL-VER Sub-TLV Capability Flags
 Registration Procedures: IETF Review
 Reference: (This document)

Bit	Description	Reference
=====	=====	=====
0	Affinity sub-TLV support.	[Affinity]
1-13	Available	
14-31	Extended header flag support.	[ExtendHeader]

Sub-Registry Name: PORT-TRILL-VER Sub-TLV Capability Flags
 Registration Procedures: IETF Review
 Reference: (This document)

Bit	Description	Reference
=====	=====	=====
0	Hello reduction support.	[ClearCorrect]
1-2	Available	
3-13	Hop-by-hop extended flag support.	[ExtendHeader]
14-31	Available	

5.5 TRILL Neighbor Record Flags

A sub-registry is created within the TRILL Parameters Registry as follows:

Sub-Registry Name: TRILL Neighbor TLV NEIGHBOR RECORD Flags
 Registration Procedures: Standards Action
 Reference: (This document)

Bit	Short Name	Description	Reference
=====	=====	=====	=====
0	Fail	Failed MTU test.	[RFC6325]
1	OOMF	Offering OOMF service.	[ClearCorrect]
2-7	-	Available.	

6. Security Considerations

For general TRILL protocol security considerations, see the TRILL base protocol standard [RFC6325].

This document raises no new security issues for IS-IS. IS-IS security may be used to secure the IS-IS messages discussed here. See [RFC5304] and [RFC5310]. Even when IS-IS authentication is used, replays of Hello packets can create denial-of-service conditions; see [RFC6039] for details. These issues are similar in scope to those discussed in Section 6.2 of [RFC6325], and the same mitigations may apply.

7. Change from RFC 6326

Non-editorial changes from [RFC6326] are summarized in the list below:

1. Additional of five sub-TLVs under the Group Address (GADDR) TLV covering VLAN labeled IPv4 and IPv6 addresses and fine-grained labeled MAC, IPv4, and IPv6 addresses. (Sections 2.1.2, 2.1.3, 2.1.4, 2.1.5, and 2.1.6).
2. Addition of the PORT-TRILL-VER sub-TLV. (Section 2.2.4)
3. Addition of the VLANs-Appointed sub-TLV. (Section 2.2.5)
4. Change the TRILL-VER sub-TLV as listed below.
 - 4.a Addition of 4 bytes of TRILL Header extended flags and capabilities supported information.
 - 4.b Require that the TRILL-VER sub-TLV appear in LSP number zero.

The above changes to TRILL-VER are backwards compatible because the [RFC6326] conformant implementations of TRILL thus far have only supported version zero and not supported any optional capabilities or extended flags, the level of support indicated by the absence of the TRILL-VER sub-TLV. Thus, if an [RFC6326] conformant implementation of TRILL rejects this sub-TLV due to the changes specified in this document, it will, at worst, decide that support of version zero and no extended flags or capabilities is indicated, which is the best an [RFC6326] conformant implementation of TRILL can do anyway. Similarly, a TRILL implementation that supports TRILL-VER as specified herein and rejects TRILL-VER sub-TLVs in an [RFC6326] conformant TRILL implementation because they are not in LSP number zero will decide that that implementation supports only version zero with no extended flag or capabilities support, which will be correct. (Section 2.3.1)

5. Clarification of the use of invalid VLAN IDs in the Appointed Forwarders sub-TLV and the Interested VLANs and Spanning Tree Roots sub-TLV. (Sections 2.2.3 and 2.3.6)
6. Addition of the Interested Labels and Spanning Tree Roots sub-TLV to indicate attachment of an IS to a fine-grained label analogous to the existing Interested VLANs and Spanning Tree Roots sub-TLV for VLANs. (Section 2.3.8)
7. Addition of the RBridge Channel Protocols sub-TLV so ISs can announce the RBridge Channel protocols they support. (Section 2.3.9)

8. Permit specification of the length of the link SNPA field in TRILL Neighbor TLVs. This change is backwards compatible because the size of 6 bytes is specially encoded as zero, the previous value of the bits in the new SIZE field. (Section 2.5)
9. Make the size of the MTU PDU Header Probe Source ID and Ack Source ID fields be the ID Length from the IS-IS PDU Header rather than the fixed value 6. (Section 3)
10. For robustness, require LSP number zero PDUs be originated as no larger than 1470 bytes but processed regardless of size. (Section 4.4)
11. Require that the originatingLSPBufferSize TLV, if present, appear in LSP number zero. (Section 4.5)
12. Create sub-registries for and specify the IANA Considerations policy for reserved and capability bits in the TRILL version sub-TLVs. (Section 5.4)
13. Addition of the distribution tree Affinity sub-TLV so ISs can request distribution tree attachments. (Section 2.3.10)
14. Add LABEL-GROUP sub-TLV analogous to the VLAN-GROUP sub-TLV. (Section 2.3.11)
15. Addition of a sub-registry for Neighbor TLV Neighbor RECORD flag bits. (Section 5.5)
16. Explicitly state that if the number of sources in a GADDR-TLV sub-TLV is zero, it indicates a listener for (*,G), that is, a listener not restricted by source. (Section 2.1)

8. Normative References

- [ISO-10589] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.
- [RFC1195] - Callon, R., "Use of OSI IS-IS for Routing in TCP/IP and Dual Environments", 1990.
- [RFC1982] - Elz, R. and R. Bush, "Serial Number Arithmetic", RFC 1982, August 1996.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4971] - Vasseur, JP. and N. Shen, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", 2007.
- [RFC5120] - Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5226] - Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5305] - Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", 2008.
- [RFC6165] - Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.
- [RFC6325] - Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "RBridges: Base Protocol Specification", RFC 6325, June 2011.
- [RFC6327] - Eastlake, D., Perlman, R., Ghanwani, A., Dutt, D., and V. Manral, "RBridges: Adjacency", RFC 6327, July 2011.
- [RFC6328] - Eastlake, D., "IANA Considerations for Network Layer Protocol Identifiers", RFC 6328, June 2011.
- [RFC6439] - Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.
- [Affinity] - draft-tissa-trill-cmt, work in progress.

[Channel] - draft-ietf-trill-rbridge-channel, work in progress.

[ClearCorrect] - draft-eastlake-trill-rbridge-clear-correct, work in progress.

[ExtendHeader] - draft-ietf-trill-rbridge-extension, work in progress.

9. Informative References

[802.1D-2004] - "IEEE Standard for Local and metropolitan area networks / Media Access Control (MAC) Bridges", 802.1D-2004, 9 June 2004.

[802.1Q-2011] - "IEEE Standard for Local and metropolitan area networks / Virtual Bridged Local Area Networks", 802.1Q-2011, 31 August 2011.

[Err2869] - RFC Errata, Errata ID 2869, RFC 6326, <http://www.rfc-editor.org>.

[RFC5304] - Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, October 2008.

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

[RFC5342] - Eastlake 3rd, D., "IANA Considerations and IETF Protocol Usage for IEEE 802 Parameters", BCP 141, RFC 5342, September 2008.

[RFC6039] - Manral, V., Bhatia, M., Jaeggli, J., and R. White, "Issues with Existing Cryptographic Protection Methods for Routing Protocols", RFC 6039, October 2010.

[RFC6326] - Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 6326, July 2011.

[MultiLevel] - draft-perlman-trill-rbridge-multilevel, work in progress.

Acknowledgements

The authors gratefully acknowledge the contributions and review by the following:

Adrian Farrel, Tissa Senevirathne, Joe Touch.

And the contributions by the following to [RFC6326]:

Mike Shand, Stewart Bryant, Dino Farinacci, Les Ginsberg, Sam Hartman, Dan Romascanu, Dave Ward, and Russ White. In particular, thanks to Mike Shand for the detailed and helpful comments.

This document was produced with raw nroff. All macros used were defined in the source files.

Authors' Addresses

Donald Eastlake
Huawei R&D USA
155 Beaver Street
Milford, MA 01757 USA

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

Anoop Ghanwani
Dell
350 Holger Way
San Jose, CA 95134 USA

Phone: +1-408-571-3500
EMail: anoop@alumni.duke.edu

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

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

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

EMail: ddutt.ietf@hobbesdutt.com

Ayan Banerjee
Cumulus Networks
1089 West Evelyn Avenue
Sunnyvale, CA 94086 USA

EMail: ayabaner@gmail.com

Copyright, Disclaimer, and Additional IPR Provisions

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

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

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

Donald Eastlake
Mingui Zhang
Huawei
Puneet Agarwal
Broadcom
Radia Perlman
Intel Labs
Dinesh Dutt
June 9, 2012

Expires: December 8, 2012

TRILL: Fine-Grained Labeling
<draft-ietf-trill-fine-labeling-01.txt>

Abstract

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

The TRILL base protocol standard supports labeling of TRILL data with up to 4K IDs. However, there are applications that require more fine-grained labeling of data. This document updates RFC 6325 by specifying extensions to the TRILL base 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 Base Protocol TRILL Data Labeling.....	5
2.3 Fine-Grained Labeling (FGL).....	5
3. Campus Wide VL versus FGL Semantic Differences.....	7
4. Coexistence with VL TRILL Switches.....	8
4.1 VL Specifiable Data Labels.....	8
5. Fine-Grained Labeling Details.....	10
5.1 Ingress Processing.....	10
5.2 Transit Processing.....	11
5.2.1 Unicast Transit Processing.....	11
5.2.2 Multi-Destination Transit Processing.....	11
5.3 Egress Processing.....	12
5.4 Appointed Forwarders and the DRB.....	13
5.5 Address Learning.....	13
5.6 ESADI Extensions.....	13
6. IS-IS Extensions.....	14
7. Comparison to Requirements.....	15
8. Allocation Considerations.....	16
8.1 IEEE Allocation Considerations.....	16
8.2 IANA Considerations.....	16
9. Security Considerations.....	17
Acknowledgements.....	17
Normative References.....	18
Informative References.....	18
Change History.....	19

1. Introduction

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

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

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

1.1 Terminology

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

DEI - Drop Eligibility Indicator [802.1Q]

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

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

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

TRILL Switch - Alternative name for an RBridge

VL - VLAN Labeling or VLAN Labeled or VLAN Label

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

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

2. Fine-Grained Labeling

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

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

2.1 Requirements

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

1. Fine-Grained

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

2. Silicon Considerations

Fine-grained labeling (FGL) should, to the extent practical, use existing features, processing, and fields that are already supported in at least some 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 FGL be required to be FGL aware. That is, it is desirable that RBridges not implementing the FGL feature and performing at least the transit forwarding function can usefully process TRILL Data frames that incorporate FGL.

4. Alternate Priority

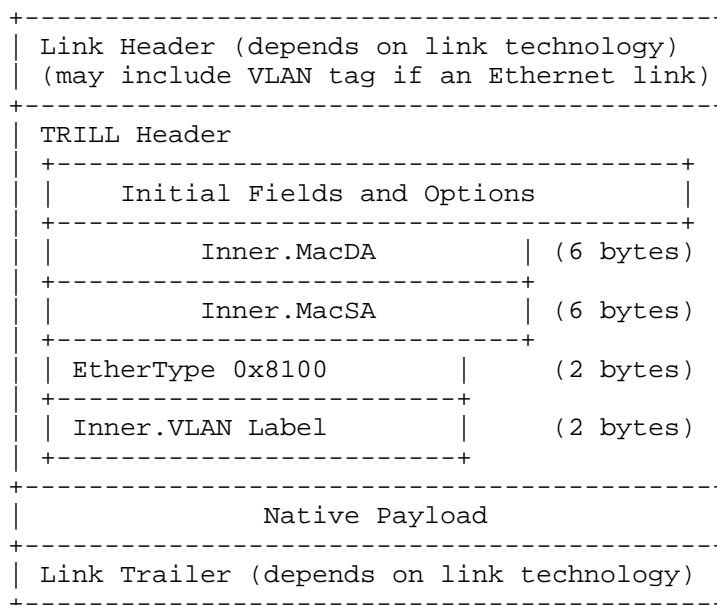
It would be desirable for an ingress TRILL Switch to be able to assign a different priority to an FGL TRILL Data frame for its

ingress-to-egress propagation from the priority of the original native frame. The original priority should be restored on egress.

2.2 Base Protocol TRILL Data Labeling

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

Currently TRILL Data frames have the VL structure shown below:

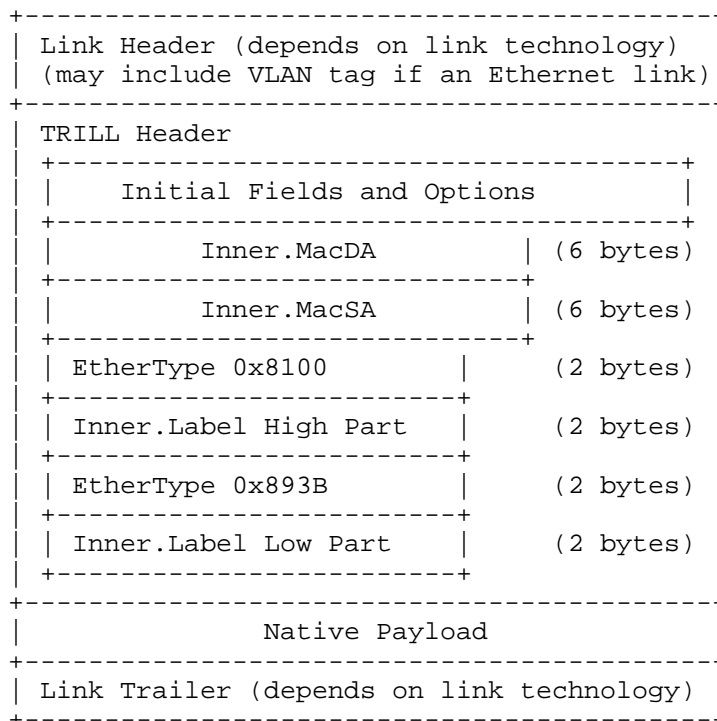


As specified in [RFC6325] the 0x8100 value is always present and is followed by the Inner.VLAN field which includes the 12-bit VLAN label.

2.3 Fine-Grained Labeling (FGL)

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

below.



The fixed format area of the TRILL Header with the Inner.Label parts and EtherType fields 0x8100 and 0x893B is mandatory for FGL frames. It is designed for backward compatibility with [RFC6325] conformant RBridges although such RBridges will only be aware of the high order 12-bits of the FGL.

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

The priority field of the Inner.Label High Part is the priority used for frame transport from ingress to egress.

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

3. Campus Wide VL versus FGL Semantic Differences

There are significant differences between the semantics across a campus for VLs and FGLs of TRILL Data frames.

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

With TRILL FGL, many things remain the same. Ports of FGL TRILL switches, at and below the EISS (Extended Internal Sublayer Service) interface, act as they do for VL R Bridges: Ethernet links between FGL TRILL switches still have only C-VLAN tagging on them and the EISS of TRILL switch ports provide a VLAN ID for an incoming frame and accepts a VLAN ID for a frame being queued for output. Appointed Forwarders [RFC6439] on a link are still appointed for a C-VLAN. The Designated VLAN for an Ethernet link is still a C-VLAN.

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

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

4. Coexistence with VL TRILL Switches

Unmodified VL RBridges will operate properly as transit TRILL switches. Transit TRILL switches look at the VL or FGL data labeling only for pruning the distribution of multi-destination frames. If an RBridge does not perform pruning, or prunes on only part of the fields in the packet, the only consequence is that multi-destination frames will use more bandwidth than necessary. VL RBridges would only look at the high order X of the (X.Y) FGL, which are in the position where a VL RBridge would expect to find a VL data label. Thus they will not be able to prune as effectively as transit FGL TRILL switches could because they will ignore the lower order half of the FGL. (Transit RBridge that fully support FGL can, of course, prune on the full FGL.)

To avoid potential problems with VL RBridges, the high order X of an (X.Y) FGL MUST NOT be zero or 0xFFFF.

It would be more serious if a VL edge RBridge, RB1, unaware of FGL, forwarded an FGL frame with FGL (X.Y) onto a link through an RB1 port configured as VL VLAN-X. VL RB1 would strip the TRILL Header only through the Inner.Label First Part and forward the packet with the Inner.Label Second Part and preceding 0x893B field still present. This might cause other problems on the link. It would also be problematic if a malicious end station could forge an apparent FGL (X.Y) frame by including extra fields in native frames ingressed by a VL edge RBridge. Therefore, it is highly desirable for all the edge RBridges to be FGL TRILL switches.

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

4.1 VL Specifiable Data Labels

It might be useful, in a particular campus with mixed VL and FGL TRILL switches, to have some end station VLANs accessible via VL edge RBridges. This is supported by reserving some number of VLANs (say the first k), to be VL-addressable. These VLANs will be specified with a VL data label, whether or not any of the edge TRILL switches attached to these end station VLANs are FGL-capable. When VL-specifiable VLANs are used in a FGL campus the upper part of an FGL MUST NOT be equal to the value of any VL-specifiable data label.

If this rule is violated, the network misconfiguration is detected by the FGL TRILL switches that will then refuse ingress to or egress from label (X.Y) while end station VLAN X connectivity is VL-specifiable as described below.

To avoid FGL frames getting pruned by VL RBridges, an FGL RBridge that ingresses to or egresses from (X.Y) MUST advertise in its LSP that it is connected to VLAN X. To avoid confusion, it is necessary to distinguish whether a TRILL switch is advertising VL-specifiable connectivity to VLAN X or just advertising such connectivity to avoid incorrect VL RBridge pruning. This is determined by whether or not the FGL RBridge advertising connectivity to VLAN X is also advertising connectivity to (X.Y) for some Y.

A VL data label X is VL-specifiable in a campus if either of the following two conditions apply:

1. A VL RBridge advertises connectivity to VLAN-X.
2. An FGL RBridge advertises connectivity to VLAN-X but does not advertise connectivity to FGL (X.Y) for any Y.

5. Fine-Grained Labeling Details

This section specifies ingress, transit, egress, and other processing of TRILL Data frames with regard to Fine-Grained Labels (FGLs). A transit or egress FGL TRILL switch detects FGL TRILL Data frames by noticing that the Inner.Label High Part is not a VL-specifiable data label (see Section 4.1).

5.1 Ingress Processing

An FGL RBridge may be configured, on one or more ports, to FGL ingress native frames. There is no change in VL ingress processing, which is the default unless a port has been configured for FGL, and no change in Appointed Forwarder logic (see Section 5.4).

FGL TRILL switches MUST support configurable per port mapping from the C-VLAN ID of a native frame, as reported by the ingress port, to an FGL. FGL TRILL switches MAY support other methods to determine the FGL of an incoming native frame, such as based on the protocol of the native frame. If the resulting label (X.Y) is such that X is a VL-specifiable data label, the ingressed frame MUST be dropped.

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

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

Use of S-tags is beyond the scope of this document but is an obvious extension.

5.2 Transit Processing

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

5.2.1 Unicast Transit Processing

There is almost no change in TRILL Data frame unicast transit processing. A transit TRILL switch forwards any unicast TRILL Data frame to the next hop towards the egress RBridge as specified in the TRILL Header. Just as transit RBridges conformant to the TRILL base protocol standard [RFC6325] do not examine the VL of unicast TRILL Data frames, transit FGL RBridges do not examine the FGL of unicast TRILL Data frames.

All transit TRILL switches, whether VL or FGL, MUST take the priority and DEI used to forward a frame from the Inner.VLAN label or the FGL Inner.Label High Part. These bits are in the same relative position for VL and FGL frames so VL RBridges will do this automatically even though they do not fully understand FGL frames.

5.2.2 Multi-Destination Transit Processing

Multi-destination TRILL Data frames are forwarded on a distribution tree selected by the ingress TRILL switch except that an FGL ingress RBridge MAY choose to TRILL unicast such a frame to all relevant egress TRILL switches if they all support FGL. The distribution trees for FGL and VL multi-destination frames are the same and are calculated as provided for in the TRILL base protocol standard [RFC6325]. There is no change in the Reverse Path Forwarding Check.

An FGL RBridge, say RB1, having an FGL multi-destination frame for label (X.Y) to forward on a distribution tree, SHOULD prune that tree based on whether there are any edge TRILL switches on a tree branch that are advertising connectivity to label (X.Y). In addition, RB1 SHOULD prune multicast frames based on reported multicast listener and multicast router attachment in (X.Y). Finally, a transit FGL RBridge MAY drop any multi-destination frame for label (X.Y) if X is VL-specifiable (see Section 4.1). "MAY" is chosen in this case to minimize the checking burden on transit TRILL switches.

To ensure that a transit VL RBridge does not falsely filter traffic for FGL (X.Y), an FGL edge RBridge reporting connectivity to FGL (X.Y) MUST report connection to VLAN X as well. Because of this, VL transit RBridges can safely apply pruning to all TRILL Data frames,

both VL and FGL, based on the reported VLAN-X connectivity of all downstream TRILL switches.

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

5.3 Egress Processing

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

If X is VL-specifiable (see Section 4.1), an FGL RBridge MUST NOT egress a frame with FGL (X.Y) but MUST drop such a frame.

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

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

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

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

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

Use of S-tags is beyond the scope of this document but is an obvious

extension.

5.4 Appointed Forwarders and the DRB

There is no change in Adjacency [RFC6327] or Appointed Forwarder logic [RFC6439] on a link regardless of whether some or all the ports on the link are for FGL R Bridges. However, if it is intended for native frames on a link in some VLAN-X to be ingressed and egressed with FGL, the Appointed Forwarder for VLAN-X for that link obviously MUST be an FGL R Bridge.

If there are FGL and VL TRILL switches connected to a link, it may be best if the priorities are configured so that the DRB is an FGL R Bridge. However, there is no inherent difficulty in a VL DRB R Bridge appointing an FGL TRILL switch connected to the link as Appointed Forwarder for whatever VLANs are appropriate.

5.5 Address Learning

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

5.6 ESADI Extensions

The TRILL ESADI (End Station Address Distribution Information) protocol is specified in [RFC6325] as optionally transmitting MAC address connection information through TRILL Data frames between participating TRILL switches over the virtual link provided by the TRILL multicast frame distribution mechanism. In [RFC6325], the VLAN to which an ESADI frame applies is indicated only by the Inner.VLAN label and no indication of that VLAN is allowed within the ESADI payload.

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

6. IS-IS Extensions

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

1. An method for a TRILL switch to announce itself in its LSP as supporting FGL.
2. A sub-TLV analogous to Interested VLANs and Spanning Tree Roots sub-TLV of the Router Capabilities TLV but indicating FGLs rather than VLANs.
3. A sub-TLV analogous to the GMAC-ADDR sub-TLV of the Group Address TLV that specifies a FGL rather than a VLAN.

See [RFC6326bis] and Section 8.2.

7. Comparison to Requirements

Comparing TRILL fine-grained labeling (FGL), as specified in this document, with the requirements given in Section 2.1, we find they are met as follows:

1. Fine-Grained: FGL provides approaching 2^{24} labels, vastly more labels than the 4K inner TRILL data labels provided in [RFC6325].
2. Silicon Considerations: Existing TRILL fast path silicon chips can, almost by definition, perform base TRILL Header insertion and removal to support ingress and egress. In addition, it is believed that most such silicon chips can also perform the native frame C-VLAN and port to fine-grained label mapping and the encoding of the fine-grained label as specified herein, as well as the inverse decoding and mapping. Some existing silicon can perform only one of these operations on a frame in the fast path and is thus not suitable to implement fast path TRILL FGL processing; however, other existing chips are believed to be able to perform both operations on the same frame in the fast path and are suitable for FGL implementation.
3. Base RBridge Compatibility: As described in Section 3, FGL is compatible with base specification (VL) RBridges [RFC6325] acting as transit TRILL switches and, as described in Section 5.4, there is no particular problem in mixing VL and FGL TRILL switches on the same link.
4. Alternate Priority: The encoding specified in Section 2.3 provides for a new priority and DEI in the Inner.Label First Part and a place to preserve the original user priority and DEI in the Second Part, so it can be restored on egress.

8. Allocation Considerations

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

8.1 IEEE Allocation Considerations

The IEEE Registration Authority has assigned EtherType 0x893B for use as the EX-TAG EtherType.

8.2 IANA Considerations

IANA is requested to allocate capability bit TBD (0 recommended) in the TRILL-VER sub-TLV capability bits [RFC6326bis] to indicate an RBridge is FGL-capable.

9. Security Considerations

See [RFC6325] for general RBridge Security Considerations.

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

Confusion between a frame with VL X and FGL (X.Y) is a potential problem:

1. A TRILL Data frame with FGL (X.Y) could be egressed to an end station in VLAN-X by a VL RBridge that is Appointed Forwarder for VLAN-X on one of its ports. This is solved by prohibiting FGL RBridges from ingressing to FGL (X.Y) if the campus is configured so that VLAN-X is VL-specifiable (see Section 4.1).
2. An end station could try to forge FGL (X.Y) frames by sending frames with an EX-TAG Y at the front to a VL RBridge port where the frame would be input as being in VLAN-X. This is solved by prohibiting egress from FGL (X.Y) while VLAN-X is VL-specifiable (see Section 4.1).

Acknowledgements

The comments and contributions of the following are gratefully acknowledged:

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

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

Normative References

- [IS-IS] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.
- [802.1Q] - IEEE 802.1, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, May 2011.
- [RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997
- [RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC6326bis] - Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-01.txt, work in progress.

Informative References

- [RFC5556] - Touch, J. and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement", RFC 5556, May 2009.
- [RFC6165] - Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.
- [RFC6327] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Dutt, D., and V. Manral, "Routing Bridges (RBridges): Adjacency", RFC 6327, July 2011
- [RFC6439] - Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.

Change History

From -00 to -01:

Update author info and make editorial changes.

Authors' Addresses

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

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

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

Email: zhangmingui@huawei.com

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

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

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

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

Dinesh G. Dutt

Email: ddutt.ietf@hobbesdutt.com

Copyright, Disclaimer, and Additional IPR Provisions

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

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

TRILL Working Group
Internet Draft
Intended status: Standard Track

Tissa Senevirathne
Les Ginsberg
CISCO

Ayan Banerjee
Consultant

Sam Aldrin
Huawei

Naveen Nimmu
Broadcom

June 17, 2012

Expires: December 2012

Multi Topology Encoding within TRILL data frames
draft-tissa-trill-mt-encode-01.txt

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/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 November 17, 2012.

Copyright Notice

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

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

Abstract

Two alternate methods of encoding Multi Topology Identifier within the TRILL data frames are presented. Methods proposed herein do not require overloading TRILL RBridge nickname to encode Multi Topology Identifier. A method that expands TRILL nickname space from 16bits to 24 bits is also presented in this draft.

Table of Contents

1. Introduction.....	3
2. Conventions used in this document.....	4
3. Multi Topology Encoding and nickname Expansion.....	4
3.1. Nickname construction.....	5
3.2. Use of Next-Hop VLAN for MT Encoding.....	6
4. Multi Topology Interoperability.....	6
4.1. Interoperability during Migration.....	6
4.2. Interoperability with RBridges with Non MT capable data plane.....	7
4.3. Interoperability with MT unaware RBRdiges.....	8
5. Backward compatibility.....	8
6. IS-IS sub-TLV definition.....	8
6.1. Nickname MSB capability.....	8
6.1.1. TRILL sub-TLVs for 24bit nicknames.....	8
6.2. MT capability.....	9
7. Security Considerations.....	9
8. Assignment Considerations.....	9
8.1. IANA Considerations.....	9
8.2. IEEE Considerations.....	10
9. References.....	10
9.1. Normative References.....	10
9.2. Informative References.....	10
10. Acknowledgments.....	11
Authors' Addresses.....	11

1. Introduction

Multi Topology is an attractive concept that allows creating different virtual topologies or overlays on top of a single physical topology. There are several important applications. Few such applications are listed below. The list below is by no means an exhaustive list and there are other applications that may utilize the MT framework.

Application specific topology (Storage topology vs. data topology)

Virtual topologies for different customers

Expansion of TRILL nickname space

Traffic Engineering

[RFC5120] presents Multi topology (MT) framework for IS-IS. MT has two important parts; IS-IS control plane extensions and Data plane encoding. [RFC5120] presents IS-IS control plane extensions. [TRILL-MT1] and [TRILL-MT2] presents required IS-IS sub-TLV definitions and the data plane encoding for TRILL.

In this document we propose methods that facilitate encoding of 16 bit Multi topology ID in to TRILL frames without reducing the effective TRILL nickname space. Additionally, the proposed scheme does not require definition of new Topology mapping sub-TLV. In other words, IS-IS control plane is similar to [RFC5120] and does not require additional complexity of mapping absolute Topology ID to abbreviated topology ID.

Methods proposed in this document encode the MT topology ID into TRILL data frames without modifying the TRILL header. Hence, MT capable, RBridges interfacing with non-MT capable RBridges can selectively not encode the proposed MT extensions on interfaces with non-MT capable RBridges. Non MT capable RBridges do not required to be in the base topology. They can be in any valid topology. Only restriction is non-MT capable RBridges can belong to a single topology only. In its IS-IS HELLO messages, RrRidges exchange its MT capability and topology information. RBridges that are not capable of supporting proposed MT extension in data plane, MUST announce itself as non MT capable, but MAY advertise its association to a topology other than the base topology by including MT extensions proposed in [RFC5120]. MT encoding capability is announced by setting the proposed MT encoding capability bit in Port TRILL Version sub-TLV [rfc6326bis]. Presence of IS-IS Multi Topology TLV

[RFC5120], indicates only the associated topology. MT encoding capability indicates RBRidges ability to support proposed data plane extensions. When MT capability is not set RBridge MUST not use the proposed data plane encoding methods, instead it must associate the announcing RBridge to the advertised topology or base topology in the absence of Multi-Topology TLV [RFC5120].

TRILL protocol, as defined in [RFC6325], defines 16bit nickname space. 16bit nickname space allows up to 65536 unique nicknames. However, it has been discussed in the working group that, more the 65536 unique names are required in certain large deployments. Possible usage of Upperbits of Nickname is also being considered for encoding Multitopology, which further reduces the available nickname space. Presented in this document is a method that allows expanding the 16bit nickname space to a 24bit nickname space, without modifying the TRILL header defined in [RFC6325].

2. Conventions used in this document

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

3. Multi Topology Encoding and nickname Expansion

[RFC6325] TRILL Base Protocol, proposes encoding scheme of TRILL frames. TRILL frames contain outer MAC Header, 802.1QTAG, TRILL header and user Data. We propose to include multi topology ID after the 802.1Q TAG. Multi Toplogy ID is preceded by Ethernet Type MT-ETHTYPE.

The expansion of nickname space requires expansion of both ingress and egress nickname spaces. We propose to encode 8bit MSB of egress nickname followed by 8bit MSB of ingress nickname in to the outer header. Figure 1 below depicts the proposed encoding.

Outer Ethernet Header:

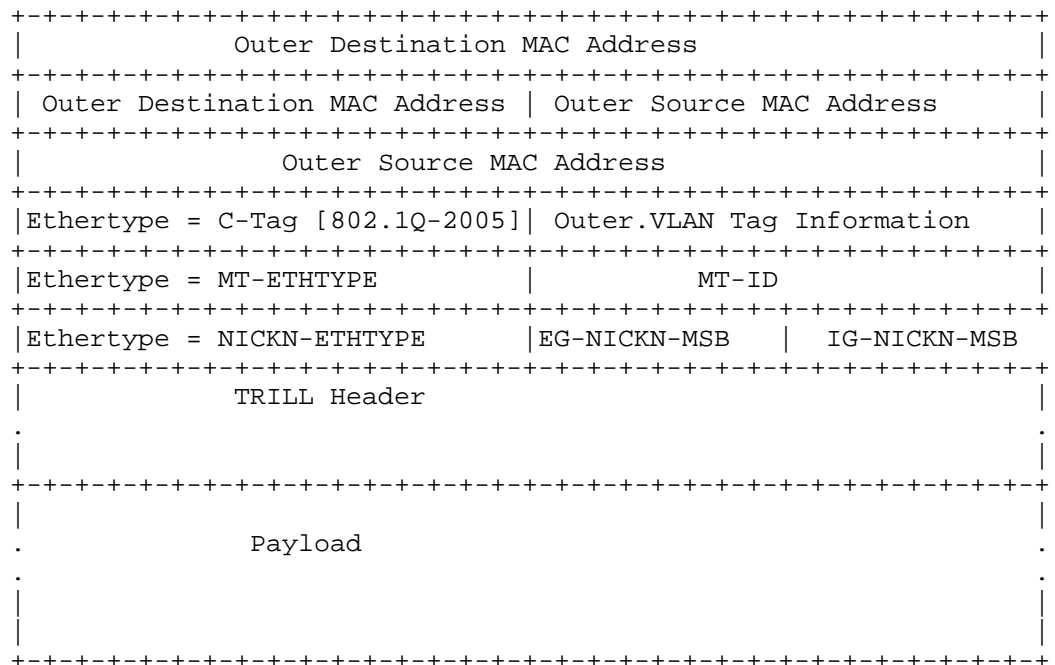


Figure 1 MT ID /NICKN MSB extentions in TRILL

Ethtype-MT : 16 bit Ethtype to encode Multi Topology ID.

MT-ID : 16bit Multi Topology ID
NICKN-Ethtype : 16 bit Ethtype to encode nickname expansion.

EG-NICKN-MSB : 8bit MSB of the Egress nickname

IG-NICKN-MSB : 8bit MSB of the Ingress nickname

3.1. Nickname construction

Each RBridge that is compatible with the proposed scheme, First check for presence of Nick-Ethtype, if present extract the EG-NICKN-MSB and IG-NICKN-MSB. EG-NICKN-MSB and IG-NICKN-MSB are then concatenated with the Egress TRILL nickname and the Ingress TRILL nickname to form 24bit nickname space. The derived nickname MUST be utilized for forwarding.

3.2. Use of Next-Hop VLAN for MT Encoding

Alternatively, Next-Hop VLAN may be utilized to encode MT ID in point to point only networks. Next Hop VLAN on TRILL outer header is independent of the inner VLAN. On Point-Pont links Next Hop VLAN is only required to be of local significance. Hence, we propose to map topologies to Next-Hop VLAN per link basis.

For sake of simplicity, we propose to map topology-id to Next-VLAN based on local policies such as configuration.

4. Multi Topology Interoperability

There are three possible scenarios of Interoperability with RBridges that are non-MT capable.

1. Interoperability During migration.
2. Interoperability with RBridges that are non-MT capable in the data plane. (i.e. Software is MT aware and supports the extensions specified here-in but, data plane is not capable of supporting the proposed encoding methods).
3. Interoperability with Rbridges that are MT unaware in both Control and data planes.

4.1. Interoperability during Migration

We recommend upgrading from the core to the edge, as depicted in the figure below. With this approach, different clusters of RBridges may belong to different topologies or to the same topology. RBridges in the core provide connectivity to RBridge clusters at the edge in a topology aware manner.

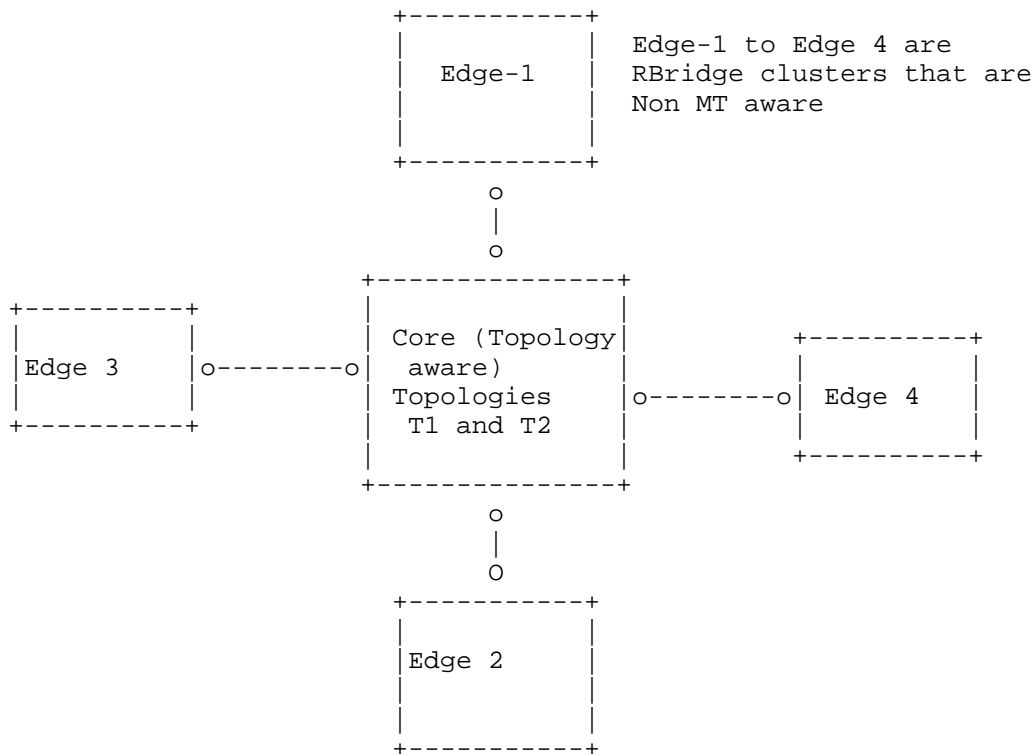


Figure 1 MT aware interconnect

In the above diagram Core may be configured to connect Edge 1 and Edge 2 to a different Topology than the topology of Edge 3 and Edge 4.

RBridges in Edge 1 - 4 are not required to be MT capable or aware. RBridges in the core associate the corresponding links to the appropriate topology.

4.2. Interoperability with RBridges with Non MT capable data plane

RBridges with Non MT capable data plane may implement MT support by dedicating a separate link for each topology.

Alternatively, RBridges, on point-point links may assign a different next hop VLAN for different topologies and derive topology ID based on VLAN. Use Next-Hop VLAN reduces the need for multiple physical

links. This method may be utilized as a permanent method for MT encoding in Point-Pont only networks.

4.3. Interoperability with MT unaware RBRdiges

MT aware RBridges identify MT unaware RBridges with either not presence of capability flags (pre RFC6326bis) or MT capability flags not being set (Section 6.2.). In such an event MT aware RBridges MUST only forward traffic related to the base topology to MT unaware RBridges. Additionally, proposed encoding MUST be removed prior to forwarding to MT unaware RBRdiges.

5. Backward compatibility

The proposed methods are encoded as part of the outer header of the TRILL frame. An RBridge that is aware of the proposed extensions when interfacing with an RBridge that is not capable of the proposed extensions MUST remove the proposed encoding from the outer header, prior to transmission of TRILL frames on those links that has RBridges that are not capable of the proposed extensions.

6. IS-IS sub-TLV definition

6.1. Nickname MSB capability

To enable auto configuration and detection of Nickname MSB capability of the peer RBRIDGE, We propose to define a flag to indicate the Nickname MSB capability.

If this capability is not present in the peer RBridge, the ETH-NICKN-MSB will be stripped out from the frame before , the frame is forwarded to the Nickname MSB unaware RBridge.

6.1.1. TRILL sub-TLVs for 24bit nicknames

TRILL standard [RFC6325] is defined for 16bit nickname space. All the associated sub-TLVs [RFC6326] are also defined to accommodate 16bit nicknames. These sub-TLVs cannot be reused to accommodate 24bit nicknames as that will break the backward compatibility. Hence, we propose to request new set of ISIS sub-TLV for following TRILL specific sub-TLVs under Router Capability TLV.

Nickname sub-TLV

Tree Identifier sub-TLV

Tree Used Identifier sub-TLV

Interested VLAN and Spanning Tree Root sub-TLV

Interested Labels and Spanning Tree Root sub-TLV

Affinity sub-TLV

6.2. MT capability

We propose to define two MT capability flags within Port TRILL Version sub-TLV.

1. MT Encoding capability
2. MT to NH-VLAN Encoding capability

MT Encoding capability flag indicates the RBridge is capable encoding MT ID using ETHTYPE-MT as defined in section 3.

MT to NH-VLAN Encoding capability flag indicates the announcing RBridge is capable of using NH-VLAN to MT ID mapping as presented in section 3.2.

When both of the flags are set RBRridge SHOULD select MT Encoding capability.

7. Security Considerations

TBD

8. Assignment Considerations

8.1. IANA Considerations

IANA is requested to allocate MT Encoding capability Flag, MT to NH-VLAN Encoding capability Flags and Nickname MSB capability flag under Port TRILL version sub-TLV.

Additionally IANA is requested to allocate new set of sub-TLV code points under Router capability TLV for the following to accommodate 24bit nickname space:

24bit Nickname sub-TLV

24bit Tree Identifier sub-TLV

24bit Tree Used Identifier sub-TLV

24bit Interested VLAN and Spanning Tree Root sub-TLV

24bit Interested Labels and Spanning Tree Root sub-TLV

24bit Affinity sub-TLV

8.2. IEEE Considerations

IEEE is requested to assign new Ether Type to represent MT-ETHTYPE defined in section 3.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RF5120] Przygienda, T. et.al , "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate System (IS-ISs)", RFC 5120, February, 2008.

9.2. Informative References

- [TRILL-MT1] Manral,V. et.al., "Multiple Topology Routing Extensions for Transparent Interconnection of Lots of Links (TRILL)", draft-manral-isis-trill-multi-topo-03, Work in Progress, December, 2011.
- [TRILL-MT2] Eastlake, D. et.al., "Multi Topology TRILL", draft-eastlake-trill-rbridge-multi-topo-02, Work in Progress, January, 2012.
- [rfc6326bis] Eastlake, D. et.al., "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-02, Work in Progress, December, 2011.

10. Acknowledgments

Special acknowledgment to Dinesh Dutt, for encouragements, support and coming up with the initial idea.

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

Authors' Addresses

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

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

Les Ginsberg
CISCO Systems
510 McCarthy Blvd,
Milpitas, CA 95035.

Email: ginsberg@cisco.com

Ayan Banerjee
Consultant

Email: ayabaner@gmail.com

Sam Aldrin
HuaWei Technologies
2330 Central Expressway
Santa Clara, CA 95951, USA

Email: aldrin.ietf@gmail.com

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

Email: naveen@broadcom.com

Transparent Interconnection of Lots of
Links Working Group
Internet-Draft
Updates: 6325 (if approved)
Intended status: Standards Track
Expires: April 25, 2013

Q. Wu
W. Hao
Huawei
October 22, 2012

LSP extension for Tree Distribution Optimization across sites
draft-wu-trill-lsp-ext-tree-distr-opt-01

Abstract

This document specifies an extension to LSP for the Rbridge in one site to advertise Global VLAN scope and associated link attribute to all the Rbridges both in the site of that Border Rbridge and the other adjacent sites in the same campus. With this extension, Rbridges can prune the distribution tree of multi-destination frames according to the scope of the VLAN and link attribute defined in this document.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 25, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. Conventions used in this document	4
3. Motivations	5
4. TLV and Sub-TLV Extensions to IS-IS for Inter-site Distribution Tree	6
4.1. Global-VLANs Sub-TLV for the Router Capability TLV	6
4.1.1. Definition of Fields in Sub-TLV	6
4.2. Link-Attributes Sub-TLV extension for extended IS reachability TLV	7
5. Use of TLV and Sub-TLV for Tree Distribution Optimization across sites	8
6. Unicast Forwarding Consideration	11
7. IANA Considerations	12
8. Security Considerations	13
9. References	14
9.1. Normative References	14
9.2. Informative References	15
Appendix A. Change Logs	16
A.1. draft-wu-trill-lsp-ext-tree-distr-opt-01	16
Authors' Addresses	17

1. Introduction

Large datacenters are often multi-site in nature and may contain a large number of Rbridges in each site. A trill Campus network may also be designed to be multilevel can be divided in to multiple IS-IS [IS-IS][RFC1195]L1 Areas interconnected by L2 backbone area. Routing between Rbridges within a IS-IS L1 area/ site is known as "Level 1 routing". Routing between IS-IS L1 areas or sites is known as "Level 2 routing". The IS-IS L1 area supports Level 1 routing and consists of Rbridges within the site and link between Rbridges within the site. The L2 backbone area supports Level 2 routing and consists of Border Rbridges and links between the Border Rbridges. Border Rbridges may participate in one or more L1 areas as Level-1 Rbridges inside each site, in addition to their role as Level 2 Rbridge across sites.

In Trill campus network, Rbridges use distribution trees to forward multi-destination frames. In case of one Trill campus network having multiple sites, the traffic associated with some distributed trees may travel between sites while the traffic associated with other distributed trees may be limited to only one site and not allowed to go across other sites. The traffic spanning across sites is also referred to as the traffic with global scope. In order to support scaling and performance of large TRILL networks in the real deployments, it is desirable to forward most of Multi-destination Trill traffic within the site and reduce the traffic that is required to span across sites within the entire TRILL campus. According to The TRILL base protocol, each distribution tree SHOULD be pruned per VLAN. When it is inevitable to construct trees that have a scope across sites throughout the TRILL campus, it is necessary to treat traffic tagged with VLAN differently based on VLAN scope and distinct the link between Rbridges in one site and link between two Border Rbridge in two sites to support large scale multi-tenants application.

This document specifies an extension to LSP for the Rbridge in one site to advertise Global VLAN scope and associated link attribute to all the Rbridges both in the site of that Border Rbridge and the other adjacent sites in the same campus. With this extension, Rbridges can prune the distribution tree of multi-destination frames according to the scope of the VLAN and link attribute defined in this document.

2. Conventions used in this document

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

3. Motivations

Distinguishing global vlan from local vlan is to increase the number of tenants by not breaking the VLAN tag size limits. E.g. one campus being divided into n sites, without distinction between global vlan and local vlan, at most support $4K$ tenants. However, if we distinguish global vlan from local vlan, suppose each site support only local vlan. Then each site support $4K$ tenants, the total number of tenants supported by one campus can be increased to $4n \cdot K$. Suppose some sites support local vlan, some sites support both local vlan and global vlan, the total number of tenants supported by one campus ($4K, 4n \cdot K$).

4. TLV and Sub-TLV Extensions to IS-IS for Inter-site Distribution Tree

This section describes data formats and code points for the TLVs and sub-TLVs added to IS-IS defined by this specification to support the multi-level TRILL or re-used from that already contained in the standard IS-IS extensions defined in [RFC6326].

4.1. Global-VLANs Sub-TLV for the Router Capability TLV

The optional Global-VLANs sub-TLV specifies the VLANs that have global scope and enable Construction of global multi-destination trees among different sites. It has the following format:

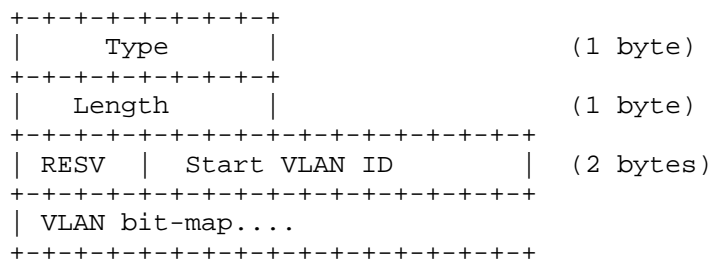


Figure 1: Report Block Structure

4.1.1. Definition of Fields in Sub-TLV

Type: 8bits

Router Capability sub-TLV type, set to TBD (GLOBAL-VLANs).

Length: 8bits

Variable, minimum 3.

RESV: 4bits

4 reserved bits that MUST be sent as zero and ignored on receipt.

Start VLAN ID:12bits

The 12-bit VLAN ID that is represented by the high order bit of the first byte of the VLAN bit-map.

VLAN bit-map:

The highest order bit indicates the VLAN equal to the start VLAN ID, the next highest bit indicates the VLAN equal to start VLAN ID + 1, continuing to the end of the VLAN bit-map field.

4.2. Link-Attributes Sub-TLV extension for extended IS reachability TLV

The link-attribute sub-TLV is carried within the TLV 22 and has a format identical to the sub-TLV format used by the Traffic Engineering Extensions for IS-IS ([RFC3784]): 1 octet of sub-type, 1 octet of length of the value field of the sub-TLV followed by the value field -- in this case, a 16 bit flags field.

The Link-attribute sub-type is 19 and the link-attribute has a length of 2 octets.

This sub-TLV is OPTIONAL and MUST appear at most once for a single IS neighbor. If a received Link State Packet (LSP) contains more than one Link-Attribute Sub-TLV, an implementation SHOULD decide to consider only the first encountered instance. The following bit is defined:

Public Link Type For TRILL(0x03) When set, this indicates that the link is public link for TRILL sites interconnection.

5. Use of TLV and Sub-TLV for Tree Distribution Optimization across sites

When the TRILL campus is divided into multiple sites, each site may have one or more Border Rbridges used to interconnect other remaining sites and form the Level 2 IS-IS Trill network. Such Level2 IS-IS Trill network can be used to construct global multi-destination tree spanning across various sites.

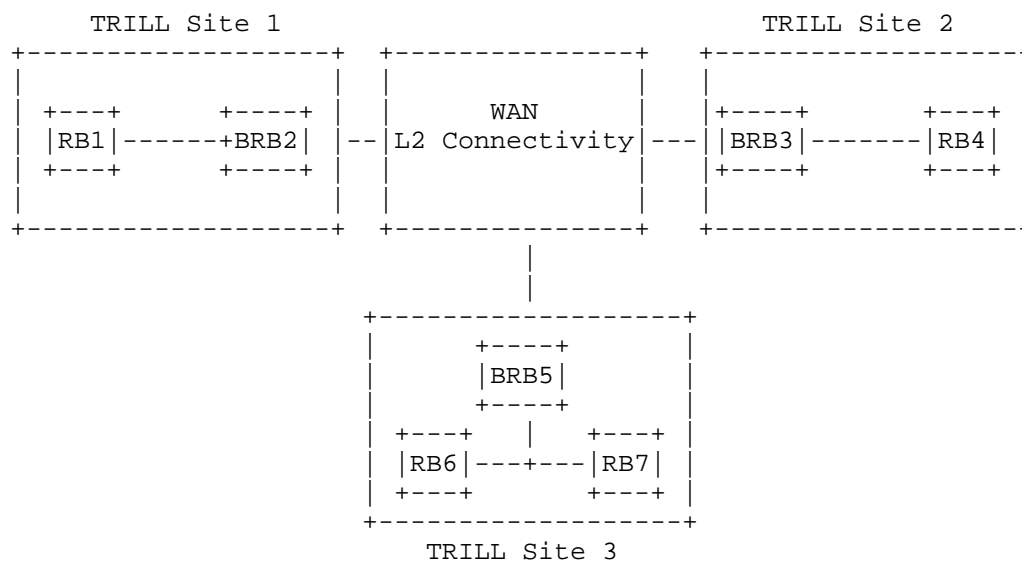


Figure 1: Example of multiple sites within one Trill Campus

In order to support scaling and performance of large TRILL networks in the real deployments, firstly, not all the links between the level 2 Rbridges need to be used to Construct global multi- destination trees. If the link between the level 2 Rbridges is allowed to construct global multi-destination trees, we can set this link attribute into "public interface for global tree construction". In this document, we reuse Link Attribute sub-TLV for the extended IS reachability TLV and allocate a new bit value inside link Attribute Sub-TLV to support indication of "public link for global tree Construction". The Border Rbridge in one site need to advertise this link attribute Sub-TLV to all the neighboring Border Rbridges in other neighboring sites and then this sub-TLV will be further forwarded to all the Rbridges in the site of each neighboring Border Rbridge. Rbridges in each site can prune the distribution tree of multi-destination frames according to such link attribute.

Secondly, not all traffic should have global scope and need to span

across sites. Since each distribution tree SHOULD be pruned per VLAN according to [RFC6325], we can specify a set of Global VLANs to identify the traffic that has global scope. In this document, we define one new sub-TLV for the Router Capability TLV, i.e., Global-VLANs Sub-TLV. This Sub-TLV can be used by Rbridges in one site to determine whether Construction of global multi-destination trees across sites is allowed. In order to achieve this, the tree root or highest priority RBridge in one site configured to know a number of appropriate VLANs as Global VLANs and announce such information to the nearest border Rbridge; Then such Border Rbridge in this site need to advertise Global VLAN Sub-TLV to all the neighboring Border Rbridges in other neighboring sites and then this sub-TLV will be further forwarded to all the Rbridges in the site of each neighboring Border Rbridge. When Global VLAN and link attribute Sub-TLV described above has been distributed to all the corresponding Rbridges in the downstream of the tree root or highest priority RBridge, Rbridges can prune the distribution tree of multi-destination frames according to the scope of the VLAN and link attribute defined in this document, eliminating branches that own link type mismatching with Distribution Tree scope identified by VLAN. If the distribution tree is local tree and has branches including a link with link attribute is set to public link for global tree construction, those branches should be eliminated. If the distribution tree is global tree and has branches containing a link with link attribute not set to public link for global tree construction, those branches also should be eliminated.

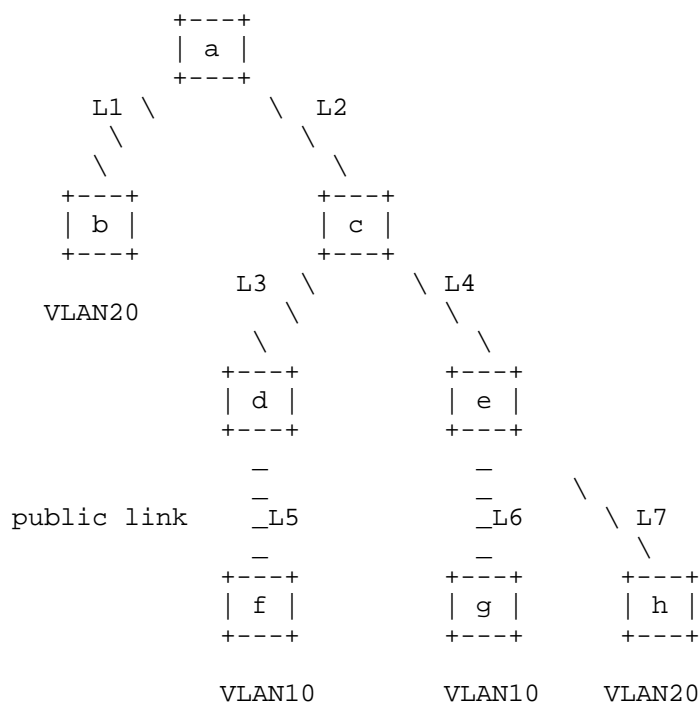


Figure 2: Distribution Tree

Take distribution tree in Figure 2 as an example, Rbridge a is root node. Rbridge f,g are leaf nodes that have end station on VLAN 10 while Rbridge b,h are another two leaf nodes and that have end station on VLAN 20. The link between Rbridge d and f is public link used across sites while the other links in the figure 2 are links owned by one single site. Assume VLAN 10 are local VLAN and VLAN 20 are Global VLAN, after distribution tree pruning is done, Rbridge c should eliminate branch that has Rbridge d and f since distribution tree is pruned based on local VLAN 10 and Link 5 in that branch is public link, which mismatch with each other.

6. Unicast Forwarding Consideration

In unicast forwarding, the MAC forwarding table for a Trill Border Rbridge is usually learned through the data plane, i.e., MAC address is learnt from received Broadcast, Unknown, Unicast, Multicast packet through distribution tree. For end stations on the local vlan, the broadcast scope is limited to one local site, the Border Rbridge only learns MAC address of locally attached end station and the forwarding path between end stations within one site can be built for unicast. For end stations on global VLAN, end stations between two sites are within the same layer 2 broadcast domain, the Border Rbridge can learn MAC address of end stations across sites and the forward path between two sites can be built as well for unicast. Therefore unicast forwarding between sites can be controlled through LSP extension we defined in this document.

If the Border Rbridge is statically configured with unicast forwarding table and the nickname of the destination Rbridge is specified as one Rbridge's nickname in other sites, the unicast packet must be forced to forward to the other sites. In this case, the Border Rbridge in other sites performs security check to the received packet. If the VLAN associated with the received packet is local VLAN and the packet is ingressed from public link across site, the packet should be discarded. If the VLAN associated with the received packet is Global VLAN, the packet should be allowed to ingress from public link across sites.

7. IANA Considerations

IANA is requested to assign a new codepoint for the Global-VLANs Sub-TLV defined in this document and carried within TLV 242.

IANA has created a registry for bit values inside the link-attributes sub-TLV called "link-attribute bit values for sub-TLV 19 of TLV 22".

This document instructs IANA to add a new bit value in the link-attribute bit values for sub-TLV 19 of TLV 22 registry as follows:

Value	Name	Reference
-----	----	-----
0x3	Public Link Type between sites	[This document]

Further values are to be allocated by the Standards Action process defined in [RFC2434], with Early Allocation (defined in [RFC4020]) permitted.

8. Security Considerations

The security considerations documented in [RFC4971][RFC5305] are applicable for the Sub-TLV extension defined in this document.

9. References

9.1. Normative References

- [IS-IS] "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002 Second Edition, 2002.
- [RFC1195] Ohta, M., "Use of OSI IS-IS for routing in TCP/IP and dual environments", December 1990.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", March 1997.
- [RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 2434, October 1998.
- [RFC3784] Smit, H., "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)", June 2004.
- [RFC4020] Kompella, K. and A. Zinin, "Early IANA Allocation of Standards Track Code Points", RFC 4020, February 2005.
- [RFC4971] Vasseur, J., "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", July 2007.
- [RFC5029] Vasseur, J. and S. Previdi, "SDP: Session Description Protocol", September 2007.
- [RFC5305] Li, T. and H. Smit, "S-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC6325] Perlman, R., Eastlake , D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC6326] Eastlake , D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 6326, July 2011.

9.2. Informative References

[TRILL-ML]

Perlman , R., Eastlake, D., Ghanwani, A., and H. Zhai,
"RBridges: Multilevel TRILL",
ID draft-perlman-trill-rbridge-multilevel-03,
October 2011.

Appendix A. Change Logs

A.1. draft-wu-trill-lsp-ext-tree-distr-opt-01

The following are the major changes to previous version draft-wu-trill-lsp-ext-tree-distr-opt-00:

- o Add one new section to discuss Unicast Forwarding.
- o Add one new section to clarify the motivation to write this draft.
- o Some other editorial changes.

Authors' Addresses

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: bill.wu@huawei.com

Weiguo Hao
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: haoweiguo@huawei.com

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: January 1, 2013

Mingui Zhang
Huawei
Tissa Senevirathne
CISCO
Janardhanan Pathangi
DELL
Ayan Banerjee
June 30, 2012

TRILL Resilient Distribution Trees
draft-zhang-trill-resilient-trees-00.txt

Abstract

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

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

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/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) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	4
1.1. Conventions used in this document	5
1.2. Terminology	5
2. Usage of Affinity TLV	5
2.1. Allocating Affinity Links	5
2.2. Distribution Tree Calculation with Affinity Links	6
3. Resilient Distribution Trees Calculation	7
3.1. Algorithm for Choosing Affinity Links	8
3.2. Affinity Links Advertisement	9
4. Resilient Distribution Trees Installation	9
4.1. Pruning the Backup Distribution Tree	10
4.2. RPF Filters Preparation	10
5. Protection Mechanisms with Resilient Distribution Trees	11
5.1. Global 1:1 Protection	11
5.2. Global 1+1 Protection	12
5.2.1. Failure Detection	12
5.2.2. Traffic Forking and Merging	13
5.3. Local Protection	13
5.3.1. Start Using Backup Distribution Tree	14
5.3.2. Duplication Suppression	14
5.3.3. An Example to Walk Through	14
5.4. Back to the Primary Distribution Tree	15
6. Security Considerations	15
7. IANA Considerations	15
8. References	16
8.1. Normative References	16
8.2. Informative References	16

Author's Addresses	18
------------------------------	----

1. Introduction

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

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

[6326bis] defines the Affinity TLV. An "Affinity Link" can be explicitly assigned to a distribution tree or trees. This offers a way to manipulate the calculation of distribution trees. With intentional assignment of Affinity Links, a backup distribution tree can be set up to protect links on a primary distribution tree. Based on this, this document proposes "Resilient Distribution Trees (RDT)" in which backup trees are installed in advance for the purpose of fast failure repair. In this way, resilience is built into the distribution tree calculation.

This document makes use of RDT to realize link protection while node protection is out of its scope. Three types of protection mechanisms are proposed. Global 1:1 protection is used to refer to the mechanism having the multicast source RBridge normally injects one multicast stream onto the primary DT. When this stream is detected to be interrupted, the source RBridge switches to the backup DT to inject subsequent multicast stream until the primary DT is recovered. Global 1+1 protection is used to refer to the mechanism having the multicast source RBridge always injects two copies of multicast streams onto the primary DT and backup DT respectively. In normal case, multicast receivers pick the stream sent along the primary DT and egress it to its local link. When a link failure interrupts the primary stream, the backup one will be picked until the primary DT is recovered. Local protection refers to the mechanism having the RBridge attached to the failed link to locally repair the failure.

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

1.1. Conventions used in this document

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

1.2. Terminology

IS-IS: Intermediate System to Intermediate System

TRILL: TRansparent Interconnection of Lots of Links

DT: Distribution Tree

RPF: Reverse Path Forwarding

RDT: Resilient Distribution Tree

SPF: Shortest Path First

SPT: Shortest Path Tree

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

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

2. Usage of Affinity TLV

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

2.1. Allocating Affinity Links

Affinity TLV explicitly assigns parents for RBridges on distribution trees. They are advertised in the Affinity TLV and recognized by each

RBridge in the campus. The originating RBridge becomes the parent and the nickname contained in the Affinity Record identifies the child, which explicitly provides an "Affinity Link" on a distribution tree or trees. The "Tree-num of roots" of the Affinity Record identify the distribution trees that adopt this Affinity Link [6326bis].

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

2.2. Distribution Tree Calculation with Affinity Links

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

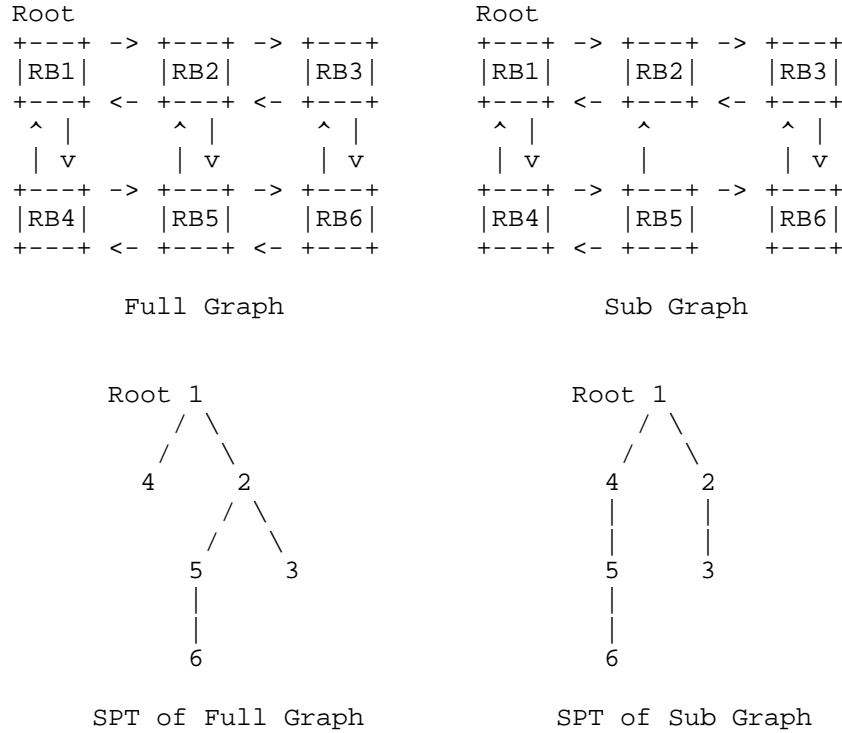


Figure 2.1: DT Calculation with the Affinity Link RB4-RB5

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

3. Resilient Distribution Trees Calculation

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

1. Failure detected through IS-IS control messages (HELLO) exchanging;
2. IS-IS flooding and each RBridge recognizes the failure;
3. Each RBridge recalculates affected distribution trees independently;

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

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

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

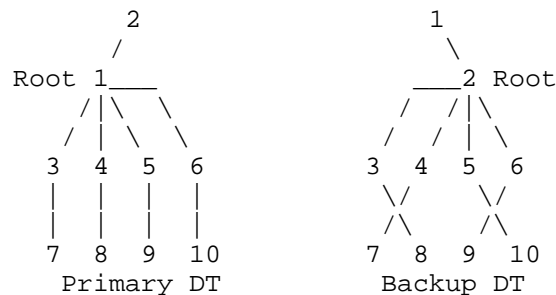


Figure 3.1: An Example of a Primary DT and its Backup DT

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

3.1. Algorithm for Choosing Affinity Links

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

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

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

3.2. Affinity Links Advertisement

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

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

4. Resilient Distribution Trees Installation

In order to reduce the service disruption time, RBridges SHOULD install backup DTs in advance. This also includes the RPF filters that need to be set up for RPF Check.

Since the backup DT is intentionally built up maximally disjoint to the primary DT, when a link fails and interrupts the ongoing multicast traffic sent along the primary DT, it is probably that the backup DT is not affected. Therefore, the backup DT installed in

advance can be used to deliver multicast frames immediately.

4.1. Pruning the Backup Distribution Tree

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

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

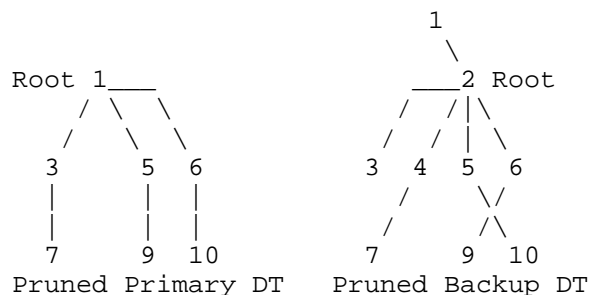


Figure 4.1: The Backup DT is Pruned Based on the Pruned Primary DT.

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

4.2. RPF Filters Preparation

RB2 includes in its LSP the information to indicate which trees RB2 might choose to ingress multicast frames [RFC6325]. When RB2 specifies the trees it might choose to ingress multicast traffic, it SHOULD include the backup DT. Other RBridges will prepare the RPF check states for both the primary DT and backup DT. When a multicast packet is sent along either the primary DT or the backup DT, it will pass the RPF Check. This works when global 1:1 protection is used.

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

5. Protection Mechanisms with Resilient Distribution Trees

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

- o The link on a TRILL distribution tree is bidirectional while the link on a distribution tree in IP/MPLS networks is unidirectional.
- o In TRILL, an multicast source node does not have to be the root of the distribution tree. It goes just the opposite in IP/MPLS networks.
- o In IP/MPLS networks, distribution trees are constructed for each multicast source node as well as their backup distribution trees. In TRILL, a small number of core distribution trees are shared among multicast groups. A backup DT does not have to share the same root as the primary DT.

Therefore TRILL needs dedicated multicast protection mechanisms.

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

5.1. Global 1:1 Protection

In the global 1:1 protection, the ingress of the multicast traffic is

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

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

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

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

5.2. Global 1+1 Protection

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

5.2.1. Failure Detection

Egress RBridges (merge points) SHOULD realize the link failure as early as possible so that failure affected egress RBridges may update

their RPF filters quickly to minimize the traffic disruption. Three options are provided as follows.

1. Egress RBridges assume a minimum known packet rate for a given data stream [MoFRR]. A failure detection timer T_d are set as the interval between two continuous packets. T_d is reinitialized each time a packet is received. If T_d expires and packets are arriving at the egress RBridge on the backup DT (within the time frame T_d), it updates the RPF filters and starts to receive packets forwarded on the backup DT.
2. With multi-destination BFD, when a link failure happens, affected egress RBridges can detect a lack of connectivity from the ingress [mBFD]. Therefore these egress RBridges are able to update their RPF filters promptly.
3. Egress RBridges can always rely on the IS-IS control plane to learn the failure and determine whether their RPF filters should be updated.

5.2.2. Traffic Forking and Merging

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

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

5.3. Local Protection

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

Since the ingress RBridge is not necessarily the root of the distribution tree in TRILL, a multicast downstream point may be not the descendants of the ingress point on the distribution tree.

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

5.3.1. Start Using Backup Distribution Tree

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

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

5.3.2. Duplication Suppression

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

5.3.3. An Example to Walk Through

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

In the normal case, multicast frames ingressed by RB7 using the pruned primary DT rooted at RB1 are being received by RB9 and RB10. When the link RB1-RB5 fails, the PLR RB1 begins to replicate and forward subsequent multicast frames using the pruned backup DT rooted at RB2. When RB2 gets the multicast frames from the link RB1-RB2, it accepts them since the RPF filter {DT=RB2, ingress=RB7, receiving

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

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

5.4. Back to the Primary Distribution Tree

After the reconvergence finishes, the new primary DT and RPF filters will be installed. The link failure will be recovered. Then the backup DT and its RPF filters will be updated as well. Before the primary DT is successfully installed and come into use, the backup DT SHOULD not be updated since repaired multicast stream may be still being forwarded on it.

For the global 1:1 protection, the ingress RBridge SHOULD switch to the new primary DT immediately after it's installed. Then the backup DT is updated.

For the global 1+1 protection, the ingress RBridge SHOULD start to use the installed primary DT, and at the same time MUST stop replicating multicast frames onto the backup DT. After the backup DT is updated, the ingress RBridge starts to replicate multicast frames onto the backup DT.

For the local protection, the ingress RBridge starts to use the new primary DT immediately after it's installed. The PLR MUST stop replicating and sending packets on the old backup DT. Then the backup DT will be updated.

6. Security Considerations

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

7. IANA Considerations

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

publication.

8. References

8.1. Normative References

- [6326bis] D. Eastlake, A. Banerjee, et al., "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-07.txt, work in Progress.
- [CMT] T. Senevirathne, J. Pathangi, et al, "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-trill-cmt-00.txt, working in progress.
- [RFC6325] R. Perlman, D. Eastlake, et al, "RBrigdes: Base Protocol Specification", RFC 6325, July 2011.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.
- [rbBFD] V. Manral, D. Eastlake, et al, "TRILL (Transparent Interconnetion of Lots of Links): Bidirectional Forwarding Detection (BFD) Support", draft-ietf-trill-rbridge-bfd-06.txt, work in progress.
- [mBFD] D. Katz, D. Ward, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-00.txt, work in progress.
- [RFC5880] D. Katz, D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

8.2. Informative References

- [mMRT] A. Atlas, R. Kebler, et al., "An Architecture for Multicast Protection Using Maximally Redundant Trees", draft-atlas-rtgwg-mrt-mc-arch-00.txt, work in progress.
- [MoFRR] A. Karan, C. Filsfils, et al., "Multicast only Fast Re-Route", draft-karan-mofrr-02.txt, work in progress.
- [RBch] D. Eastlake, V. Manral, et al, "TRILL: RBridge Channel

Support", draft-ietf-trill-rbridge-channel-06.txt, work in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies Co.,Ltd
Huawei Building, No.156 Beiqing Rd.
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

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

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

Janardhanan Pathangi
Dell/Force10 Networks
Olympia Technology Park,
Guindy Chennai 600 032

Phone: +91 44 4220 8400
Email: Pathangi_Janardhanan@Dell.com

Ayan Banerjee

Email: ayabaner@gmail.com