TRILL working group   L. Dunbar
Internet Draft   D. Eastlake
Intended status: Standard Track   Huawei
Expires: Sept 2012   Radia Perlman
                                                                 Intel
                                                          I. Gashinsky
                                                                 Yahoo
                                                      October 26, 2011

                      Directory Assisted RBridge Edge
                draft-dunbar-trill-directory-assisted-edge-03.txt

Status of this Memo

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

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

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

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/ietf/1id-abstracts.txt

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

   This Internet-Draft will expire on April 26, 2009.

Copyright Notice

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

Abstract

RBridge edge nodes currently learn the mapping between MAC addresses
and their corresponding RBridge edge nodes by observing the data
packets traversed through. When ingress RBridge receives a data
packet with its destination address (MAC&VLAN) unknown, the data
packet is flooded across RBridge domain. When there are more than
one RBridge ports connected to one bridged LAN, only one of them can
be designated as AF port for forwarding/receiving traffic for each
LAN, the rest have to be blocked for that LAN.

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

Conventions used in this document

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

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
document are to be interpreted as described in RFC-2119 0.

Table of Contents

1. Introduction

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

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

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

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

2. Terminology

   AF   Appointed Forwarder RBridge port

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

   DA: Destination Address

   DC: Data Center

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

   FDB: Filtering Database for Bridge or Layer 2 switch

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

SA: Source Address

STP: Spanning Tree Protocol

RSTP: Rapid Spanning Tree Protocol

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

VM: Virtual Machines

3. Impact on RBridge domain of massive number of hosts in Data Center

It is common for Data Center networks to have multiple tiers of
switches, e.g. one or two Access Switches for each server rack
(ToR), aggregation switches for some rows (or EoR switches), and
some core switches to interconnect the aggregation switches. Many
aggregation switches deployed in data centers are high port density
switches. It is not uncommon to see aggregation switches
interconnecting hundreds of ToR switches.

```
                +-------+           +------+
               +/------+ |         +/-----+ |
               | Aggr11| + -----  |AggrN1|  +      EoR Switches
               +---+---+/          +------+/
                /     \             /      \
               /       \           /        \
         +---+     +---+     +---+      +---+
         |T11|... |T1x|     |T21| ?  |T2y|  ToR switches
         +---+     +---+     +---+      +---+
           |         |         |          |
         +-|-+     +-|-+     +-|-+      +-|-+
         |   |... |   |     |   | ..  |   |
         +---+     +---+     +---+      +---+      Server racks
         |   |... |   |     |   | ..  |   |
         +---+     +---+     +---+      +---+
         |   |... |   |     |   | ..  |   |
         +---+     +---+     +---+      +---+
```
            Figure 1: Typical Data Center Network Design

When TRILL is deployed in a data center with large number of hosts,
with the possibility of hosts in one subnet/VLAN being placed under

multiple edge RBridges and each edge RBridge having hosts from
different subnets/VLANs, the following problems will occur:

-   Unnecessary filling of slots in MAC table of edge RBridges, due
    to edge RBridge receiving broadcast traffic (ARP/ND
    broadcast/multicast) from hosts under other edge RBridges that
    are not actually communicating with any hosts attached to the
    RBridge whose table is being unnecessarily filled.

-   Some edge RBridge ports being blocked for user traffic when
    there are more than one RBridge ports connected to one bridged
    LAN. When there are multiple RBridge ports connected to a
    bridged LAN, only one, i.e. the AF port, can forward/receive
    traffic for that bridged LAN which is normally equivalent to a
    VLAN, the rest have to be blocked for forwarding/receiving
    traffic for that VLAN. When a rack has dual uplinks to two
    different ToR switches, i.e. RBridge Edges, (which is very
    common), some links can't be fully utilized.

-   Packets being flooded across RBridge domain when their DAs are
    not in ingress RBridge's cache.

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

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

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

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

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

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

  But in this scenario, there will be multiple RBridge edge ports
  connected to one bridged LAN, which requires only one of them
  being designated as Appointed Forwarder (AF port) for
  forwarding native traffic across RBridge domain for that VLAN,
  while other ports/links being blocked for native frames in that
  VLAN. There is also possibility of loops on the bridged LAN
  attached to RBridge edge ports unless STP/RSTP is running.
  Running traditional Layer 2 STP/RSTP on the bridged LAN in this
  environment may be overkill because the topology among the ToR
  switches and aggregation switches is very simple.

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

4. Directory Assisted RBridge Edge in DC Environment

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

   Avoiding unknown DA flooding to RBridge domain is especially
   valuable in data center environment because there is higher chance
   of an RBridge edge receiving packets with unknown DA and
   broadcast/multicast messages due to VM migration and servers being
   loaded with different applications. When a VM is moved to a new

location or a server is loaded with a new application with different
IP/MAC addresses, it is more likely that the DA of data packets sent
out from those hosts are unknown to their attached RBridge edges.
In addition, gratuitous ARP (IPv4) or Unsolicited Neighbor
Advertisement (IPv6) sent out from those newly migrated or activated
hosts have to be flooded to other RBridge edges which have hosts in
the same subnets.

The benefits of using directory assistance include:

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

  When multiple RBridge edge ports are connected via bridged LAN
  to hosts (servers/VMs), a directory assisted RBridge edge can
  simply drop frames with an unknown DA. It won't need to flood
  those data frames across RBridge domain. Therefore, there is no
  need to designate one Appointed Forwarder among all the RBridge
  Edge ports connected to a bridge LAN, which means that all
  RBridge ports can forward/receive traffic.

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

  When an RBridge receives a TRILL frame whose destination
  Nickname matches with its own, the normal procedure is for the
  RBridge to decapsulate the TRILL header and forward the
  decapsulated Ethernet frame to its directly attached bridged
  LAN. If the destination MAC is unknown, the decapsulated
  Ethernet frame is flooded in the LAN. With directory
  assistance, the RBridge edge can determine if DA in a frame
  matches with any hosts attached via the bridged LAN. Therefore,
  frames can be discarded if their DAs do not match.

- Reduce the amount of MAC&VLAN <-> RBridgeEdge mapping
  maintained by RBridge edge. There is no need for an RBridge
  edge to keep the MAC entries for hosts which don't communicate
  with hosts attached to the RBridge edge.

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

4.1. Push Model

Under this model, Directory Server(s) push down the MAC&VLAN <->
RBridgeEdge mapping for all the hosts which might communicate with
hosts attached to an RBridge edge node. The mapping entry to be
pushed down could leverage the gratuitous ARP reply with extended
fields showing the edge RBridge's name, as shown in Table 2. Using
Table 2 requires one entry per host. When directory pushes down the
entire mapping to an edge RBridge for the very first time, there
usually are many entries. To minimize the number of entries pushed
down, summarization should be considered, e.g. with one edge RBridge
Nickname being associated with all attached hosts' MAC addresses and
VLANs as shown below:

```
   +------------+-------+------------------------------+
   | Nickname1  |VID-1  | MAC1, MAC2, ..MACn           |
   |            |------ +------------------------------+
   |            |VID-2  | MAC1, MAC2, ..MACn           |
   |            |------ +------------------------------+
   |            |...    | MAC1, MAC2, ...MACn          |
   +------------+-------+----------------------------------+
   | Nickname2  |VID-1  | MAC1, MAC2, ...MACn              |
   |            |------- +---------------------------------+
   |            |VID-2  | MAC1, MAC2, ...MACn              |
   |            |------- +---------------------------------+
   |            |...    | MAC1, MAC2, .. MACn              |
   +------------+-------+---------------------------------+
   | -------    |------- +---------------------------------+
   |            |....   | MAC1, MAC2, ... MACn             |
   +------------+-------+---------------------------------+
```
         Table 1: Summarized table pushed down from directory

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

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

It may not be necessary for every RBridge edge to get the entire
mapping table for all the hosts in a data center. There are many
ways to narrow the full set down to a smaller set of remote hosts
which communicate with hosts attached to an RBridge edge. A simple
approach of only pushing down the mapping for the VLANs which have

active hosts under an RBridge edge can reduce the number of mapping entries pushed down.

However, it is inevitable that RBridge edge's MAC&VLAN<->RBridgeEdge mapping table will have more entries than they really need under the Push Model. When hosts attached to one RBridge Edge rarely communicate with hosts attached to different RBridge edges even though they are on the same VLAN, the normal process of RBridge edge's unknown DA flooding, learning and cache aging would have removed those MAC&VLAN entries from the RBridge's cache. But it can be difficult for Directory Servers to predict the communication patterns among hosts within one VLAN. Therefore, it is likely that the Directory Servers will push down all the MAC&VLAN entries if there are hosts in the VLAN being attached to the RBridge Edge.

4.2. Pull model:

Under this model, ''RBridge'' pulls the MAC&VLAN<->RBridge mapping entry from the directory server when needed. RBridge edge node can simply intercept all ARP/ND requests and frames with unknown DA, and forward them to the Directory Server(s) that has the information on where each host is located.

The reply from the Directory Server can be the standard ARP/ND reply with an extra field showing the RBridge egress node's Nickname, as depicted in Table 2. RBridge ingress node can cache the mapping.

If RBridge edge node receives a data packet with unknown MAC-DA, it can query the directory server. If there is no response from the directory server, the RBridge edge node can drop the packet.

One advantage of the Pull Model is that RBridge edge can age out MAC&VLAN entries if they haven't been used for a certain period of time. Therefore, each RBridge edge will only keep the entries which are frequently used, i.e. mapping table size can be smaller. RBridge edge would query the Directory Server(s) for unknown DAs in data frames or ARP/ND and cache the response. When hosts attached to one RBridge Edge rarely communicate with hosts attached to different RBridge edges even though they are on the same VLAN, the corresponding MAC&VLAN entries would be aged out from the RBridge's cache.

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

```
         0                   1                   2                   3
         0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        | Hardware Type                 |          protocol Type        |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        | HLEN          | PLEN          |        Operation              |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |              Sender Hardware Address (MAC)                    |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |Sender Hardware Address' cont  | Sender Protocol Address (IP)  |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |Sender Protocol Address' cont  | Target Hardware Address (MAC) |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |             Target Hardware Address' cont (MAC)              |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |             Target Protocol Address (IP)                     |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      ->|             Ingress RBridge's Nickname                       |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      ->|Ingress RBridge's Nickname ext |   Egress RBridge's Nickname   |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      ->|            Egress RBridge's Nickname extension               |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
            Table 2: Extended fields added to standard ARP reply
```

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

5. Conclusion and Recommendation

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

   Therefore, we suggest TRILL consider directory assisted
   approach(es). This draft only introduces the basic concept of using
   directory assisted approach for RBridge edge nodes to learn the
   MAC&VLAN<->RBridgeEdge mapping. More complete mechanisms will be
   developed after the working group reaches some level of consensus.

6. Manageability Considerations

   TBD.

7. Security Considerations

   TBD.

8. IANA Considerations

   TBD

9. Acknowledgments

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

10. References

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

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

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

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

Authors' Addresses

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

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

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

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

users of this specification can be obtained from the IETF on-line
IPR repository at http://www.ietf.org/ipr

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

Disclaimer of Validity

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

Acknowledgment

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

TRILL working group   L. Dunbar
Internet Draft   D. Eastlake
Intended status: Standard Track   Huawei
Expires: Sept 2012   Radia Perlman
                                                              Intel
                                                       I. Gashinsky
                                                              Yahoo
                                                   October 26, 2011

                  Directory Assisted TRILL Encapsulation
              draft-dunbar-trill-directory-assisted-encap-01.txt

Status of this Memo

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

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

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

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/ietf/1id-abstracts.txt

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

   This Internet-Draft will expire on April 26, 2009.

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

Abstract

This draft describes how data center network can benefit from non-
RBridge nodes performing TRILL encapsulation and how directory
service can assist a non-RBridge node to encapsulate TRILL header.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
document are to be interpreted as described in RFC-2119 0.

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

Table of Contents

1. Introduction

It is no longer uncommon for a data center to have thousands of
server racks. Those thousands of server racks could be connected by
multiple groups of aggregation switches, with each group connecting
hundreds of ToR switches. For servers supporting virtualization,
there is typically a virtual switch embedded in each physical
server.

When TRILL is deployed in those data centers, there are issues no matter where the RBridge domain boundary starts. If RBridge domain boundary starts at aggregation switch level, the RBridge's IS-IS routing scales well, but there are problems with allowing only one (AF port) of multiple ports connected to a bridged LAN for forwarding traffic and requiring each RBridge edge to maintain a very large table of MAC&VLAN<-> RBridgeEdge mapping. If the RBridge domain boundary starts closer to hosts, e.g. at the virtual switches on servers, the number of MAC&VLAN<->Edge mapping is much smaller because each virtual switch only needs to maintain the mapping for remote hosts which actually communicate with the embedded VMs. But then, the number of nodes in RBridge IS-IS domain is very large, making it not scale well especially on aggregation switches which need to advertise link state over hundreds of ports.

[RBridge-directory] introduces a method for RBridge edge to get MAC&VLAN<->RBridgeEdge mapping from a directory service in data center environment instead of flooding unknown DAs across TRILL domain. When directory is used, any node, even non-RBridge node, can perform the TRILL encapsulation. This draft is to demonstrate the benefits of non-RBridge nodes performing TRILL encapsulation.

2. Terminology

AF   Appointed Forwarder RBridge port

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

DA: Destination Address

DC: Data Center

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

FDB: Filtering Database for Bridge or Layer 2 switch

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

SA: Source Address

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

VM: Virtual Machines

3. Directory Assistance to Non-RBridge

With directory assistance [RBridge-Directory], a non-RBridge can
determine if a packet should be forwarded across the RBridge domain.
Suppose the RBridge domain boundary starts at network switches (i.e.
not virtual switches embedded on servers), a directory can assist
Virtual Switches embedded on servers to encapsulate proper TRILL
header by providing the information of the RBridge edge to which the
target is attached.

```
   \              +-------+          +------+ TRILL Domain/
    \            +/------+ |        +/-----+ |        /
     \           | Aggr11| + ----- |AggrN1| +       /
      \          +---+---+/        +------+/ \      /
       \         /     \          /      \      /
        \       /       \        /        \    /
         \ +---+   +---+    +---+    +---+   /
          \- |T11|... |T1x|    |T21| .. |T2y|---
            +---+   +---+    +---+    +---+
             |       |        |        |
           +-|-+   +-|-+    +-|-+    +-|-+
           |   |...| V |    | V | .. | V |<-Virtual Switch
           +---+   +---+    +---+    +---+
           |   |...| V |    | V | .. | V |
           +---+   +---+    +---+    +---+
           |   |...| V |    | V | .. | V |
           +---+   +---+    +---+    +---+
```
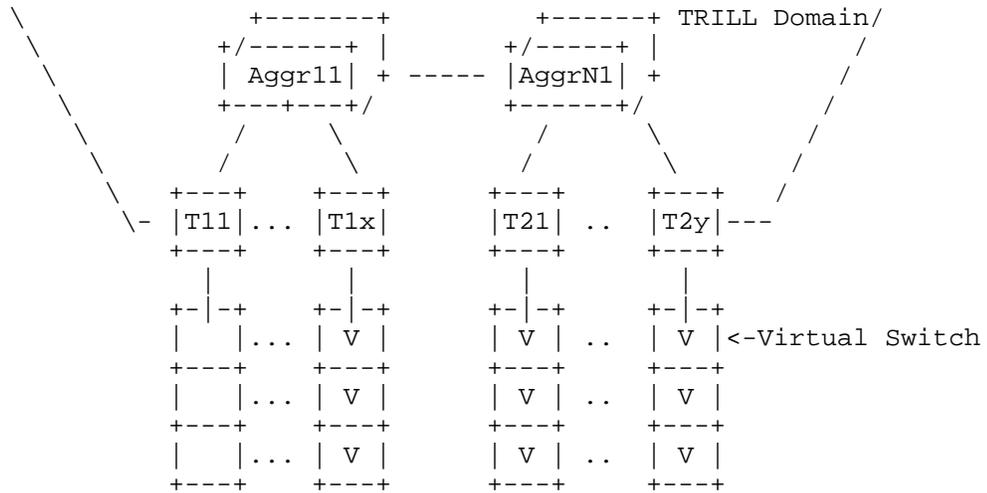   Figure 1: TRILL domain in typical Data Center Network

When a TRILL encapsulated data packet reaches an RBridge, the
RBridge can simply forward the pre-encapsulated packet to the
RBridge whose nickname is in the DA field of the TRILL header. By
doing this, no ingress RBridge will receive a native frame with
unknown DA, therefore, it won't need to flood received data packets
to all other ports. That means there is no need to worry about AF
ports and all RBridge edge ports connected to one bridged LAN can
receive and forward pre-encapsulated traffic, which greatly improves
the overall network utilization.

([RBridge] Section 4.6.2 Bullet 8 specifies that an RBridge port can
be configured to accept TRILL encapsulated frames from a neighbor
that is not an RBridge.)

When data frames do not need to traverse RBridge domain, they are switched by all nodes/ports per IEEE802.1Q and RBridge edge will not encapsulate and forward native Ethernet frames across RBridge domain.

When a pre-encapsulated TRILL frame arrives at an RBridge whose nickname matches with the destination nickname in the TRILL header, the processing is exactly same as normal, i.e. it decapsulates the native frame from the received TRILL frame and forwards the decapsulated Ethernet frame to the host attached to its edge ports.

We call a node which only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing a ''TRILL Encapsulating node'' or ''Simplified RBridge''. The TRILL Encapsulating Node gets the MAC&VLAN<->RBridgeEdge mapping table pushed down or pulled from directory servers [RBridge-directory]. Upon receiving a native Ethernet frame, the TRILL Encapsulating Node checks the MAC&VLAN<->RBridgeEdge mapping table, and perform the corresponding TRILL encapsulation if the entry is found in the mapping table. If the destination address and VLAN of the received Ethernet frame doesn't exist in the mapping table, the Ethernet frame is forwarded per IEEE802.1Q.

```
      +------------+--------+---------+---------+--+-------+----+
      |OuterEtherHd|TRILL HD| InnerDA | InnerSA |..|Payload| FCS|
      +------------+--------+---------+---------+--+-------+----+
           ^
           |                 |<Inner Ether Header>  |
           |
           |
           |          +-------+  TRILL    +------+
           |          | R1    |-----------|  R2  | Decapsulate TRILL
           |          +---+---+  domain   +------+ header
           |              |                   |
      +----------|        |                   |
           |              |                   |
           +-----+        +-----+
   Non-RBridge node:|T12 |                | T22 |
   Encapsulate TRILL+-----+                +-----+
   Header for data
   Frames to traverse
   TRILL domain.
```

4. Source Nickname in Frames Encapsulated by Non-RBridge Nodes

   The TRILL header includes a Source RBridge's Nickname (ingress) and
   Destination RBridge's Nickname (egress). When a TRILL header is
   added by a non-RBridge node, using the Ingress RBridge edge node's
   nickname in the source address field will make the ingress RBridge
   node receive TRILL frames with its own nickname in the frames'
   source address field, which can be confusing.

   To avoid confusion of edge RBridges receiving TRILL encapsulated
   frames with their own nickname in the frames' source address field
   from neighboring non-RBridge nodes, a new nickname can be given to
   an RBridge edge node, e.g. Phantom Nickname, to represent all the
   TRILL Encapsulating Nodes attached to the RBridge edge node.

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

5. Conclusion and Recommendation

   As the number of hosts in data center gets large, the number of
   switches interconnecting them could increase to a point that TRILL
   no longer scales well. The situation will get worse as hypervisors
   on servers are equipped with virtual switches. Therefore, we
   suggest TRILL consider directory assisted non-RBridge encapsulation
   approach. The non-RBridge encapsulation approach is especially
   useful when there are many servers in a data center equipped with
   hypervisor-based virtual switches because it is relatively easy for
   virtual switches, which are usually software based, to get directory
   assistance and perform network address encapsulation.

6. Manageability Considerations

   TBD.

7. Security Considerations

   TBD.

8. IANA Considerations

   TBD

9. Acknowledgments

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

10. References

    [RBridge-Directory] Dunbar, et, al ''Directory Assisted RBridge
    Edge'', <draft-dunbar-trill-directory-assisted-edge-02.txt>, Oct,
    2011

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

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

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

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

Authors' Addresses

    Linda Dunbar
    Huawei Technologies
    1700 Alma Drive, Suite 500
    Plano, TX 75075, USA
    Phone: (972) 543 5849
    Email: ldunbar@huawei.com

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

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

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

Disclaimer of Validity

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

Acknowledgment

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

Network Working Group   Donald Eastlake
INTERNET-DRAFT   Huawei
Intended status: Proposed Standard   Ayan Banerjee
Obsoletes: 6326   Dinesh Dutt
                                                             Cisco
                                                   Anoop Ghanwani
                                                           Brocade
                                                     Radia Perlman
                                                             Intel
Expires: April 30, 2012   October 31, 2011

         Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS
                  <draft-eastlake-isis-rfc6326bis-01.txt>

Abstract

   The IETF TRILL (Transparent Interconnection of Lots of Links)
   standard provides optimal pair-wise data frame forwarding without
   configuration in multi-hop networks with arbitrary topology, 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

Table of Contents

1. Introduction

   The IETF has standardized the TRILL (Transparent Interconnection of
   Lots of Links) protocol [RFC6325] [RFC6327], which provides
   transparent forwarding in multihop networks with arbitrary topology
   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) and can incrementally replace IEEE 802.1 customer bridges.

   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]. The main changes from [RFC6326]
   are summarized below and listed in more detail in Section 7.

   1. Fix the one reported errata in [RFC6326].

   2. Addition of multicast group announcements by IPv4 and IPv6
      address.

   3. Addition of control plane support for 24-bit TRILL Data frame
      labels that are in some ways analogous to VLANs.

   4. Addition of facilities for announcing capabilities supported.

   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

   customer bridge - A device conformant with [802.1D-2004] or
       [802.1Q-2011] that is not offering any Provider bridging
       facilities.

    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
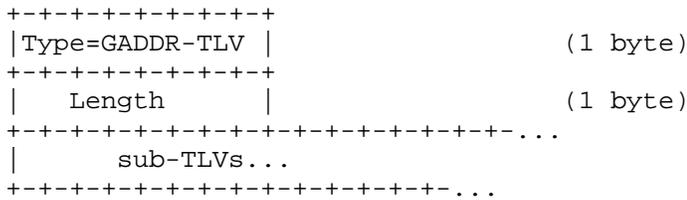   document are to be interpreted as described in [RFC2119].

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 standard. 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)         |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
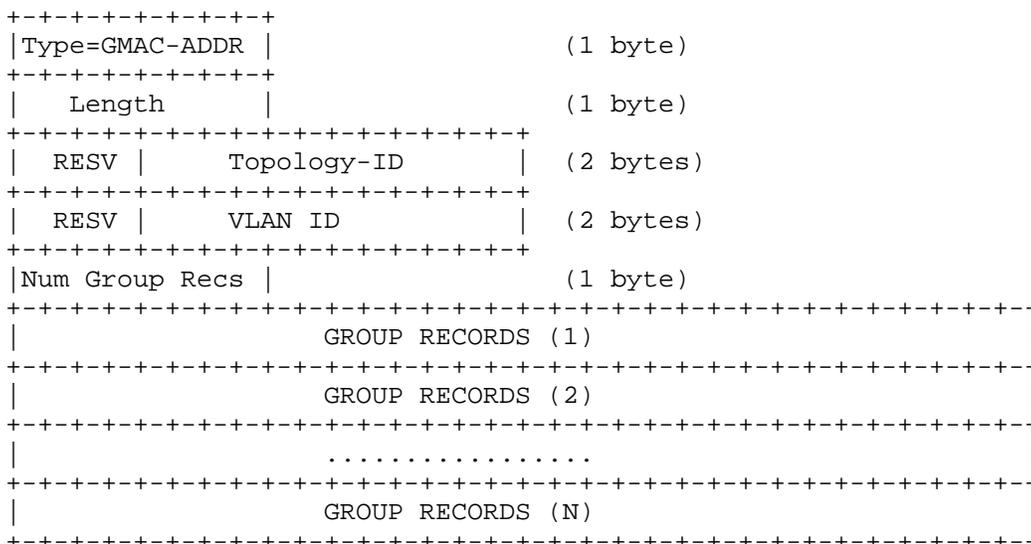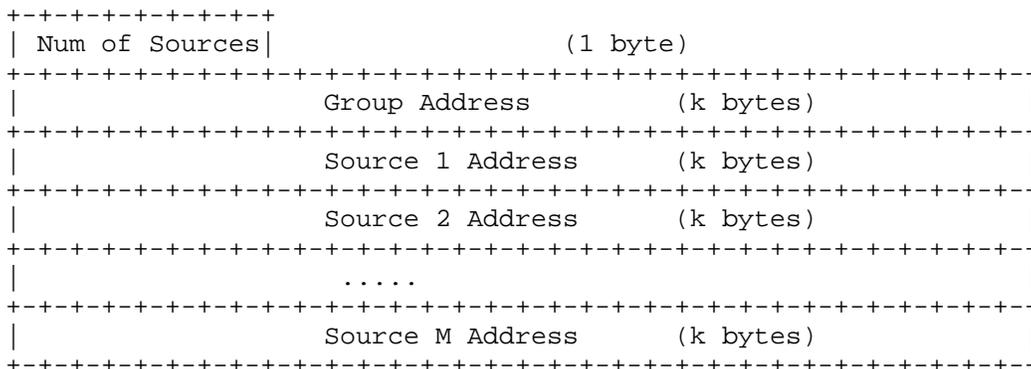
   o Type: GADDR sub-TLV type, set to 1 (GMAC-ADDR).

   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. 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.

2.1.2 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. 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.
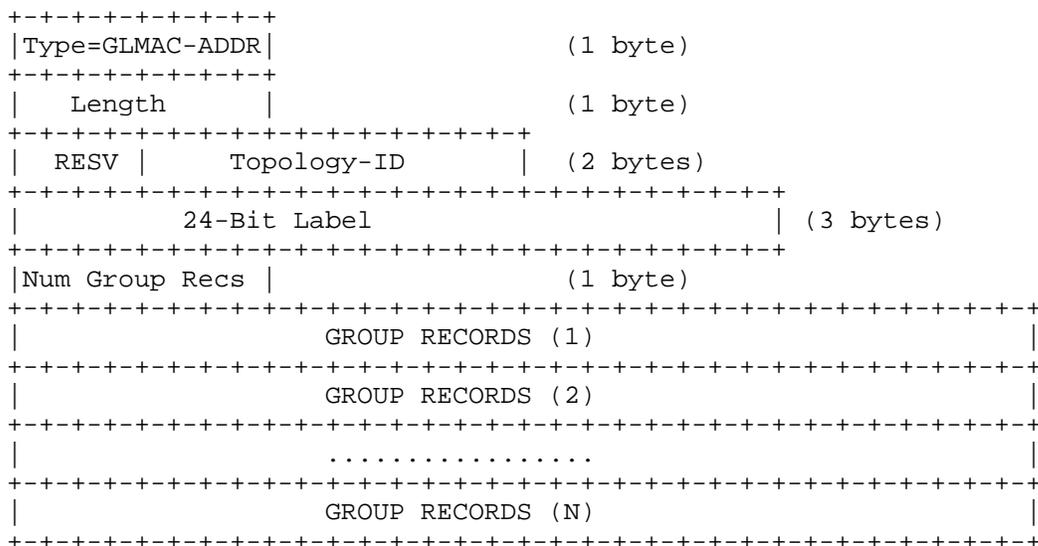
2.1.3 Group IPv6 Address sub-TLV

   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. 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-
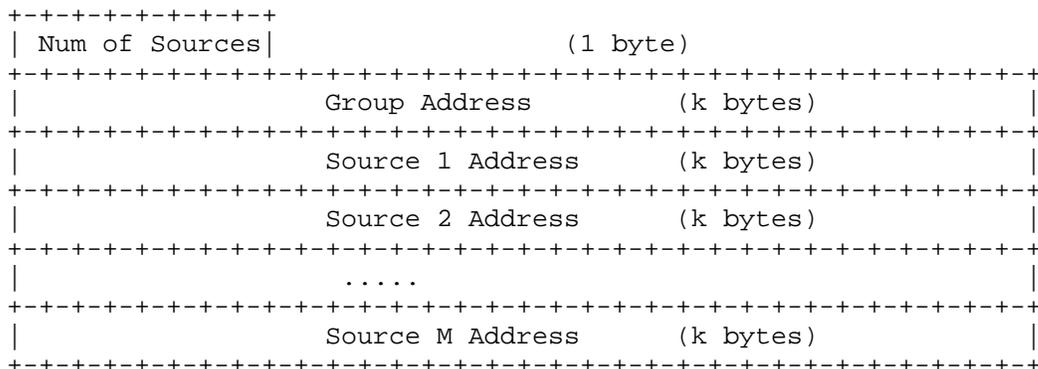   TLV of another instance of the Group Address TLV.

   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 12-bit VLAN-ID. The Group Labeled MAC
   Address sub-TLV, below, extends this to a 24-bit fine-grained label.

```
   +-+-+-+-+-+-+-+-+-+
   |Type=GLMAC-ADDR|                      (1 byte)
   +-+-+-+-+-+-+-+-+-+
   |    Length     |                      (1 byte)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  RESV |     Topology-ID     |        (2 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |            24-Bit Label                 |  (3 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |Num Group Recs |                      (1 byte)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    GROUP RECORDS (1)                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    GROUP RECORDS (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 24-bit fine-grained label identifier for
   all subsequent MAC addresses in this sub-TLV, or the value zero if

no label is specified.

o Number of Group Records: A 1-byte unsigned integer that is the
   number of group records in this sub-TLV.

o Group Record: Each group record carries the number of sources. It
   then has a 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 24-bit 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. 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 24-bit 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. 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 remainder of this section
   specifies 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.

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 [RFCaf]. 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 AppointedFwrdrs 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.

   o RESV: 4 bits that MUST be sent as zero and ignored on receipt.

   o Start.VLAN, End.VLAN: These fields are the VLAN IDs of the
      appointment range, inclusive. To specify a single VLAN, the VLAN's
      ID appears as both the start and end VLAN. As specified in
      [RFCaf], appointing an IS forwarder on a port for a VLAN not
      enabled on that port has no effect.

   An IS's nickname may occur as appointed forwarder for multiple VLAN
   ranges by occurrences of this sub-TLV within the same or different MT
   Port Capability TLVs within an IIH PDU. See [RFCaf].

2.2.4 Port TRILL Version Sub-TLV

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

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

   o Type: MT-PORT-CAP sub-TLV type, set to TBD (PORT-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 3 through 13 indicate
      that the corresponding TRILL Header hop-by-hop extended flags
      [ExtendHeader] are supported. Bits 0 through 2 and 14 to 31 are
      reserved to indicate support of optional capabilities. A one bit
      indicates that the flag or capability is supported by the
      originating IS. Bits in this field MUST be set to zero except as
      permitted for a capabilities 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 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
TRILL-PORT-VER sub-TLV and an IS, if it is adjacent to the
originating IS of TRILL-VER sub-TLV(s) uses the TRILL version it
received in TRILL-PORT-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.

   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 VLANs
   for which the originating IS is declaring it is appointed forwarder
   on the port on which the enclosing IISH was sent is 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)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+-+
                     1 1 1 1 1 1 1 1 1   3 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 TRILL-PORT-VER sub-TLV (Section 2.2.4) so they cannot
conflict. However, the TRILL version is the same as that in the
TRILL-PORT-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
TRILL-PORT-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)
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...+-+-+
```
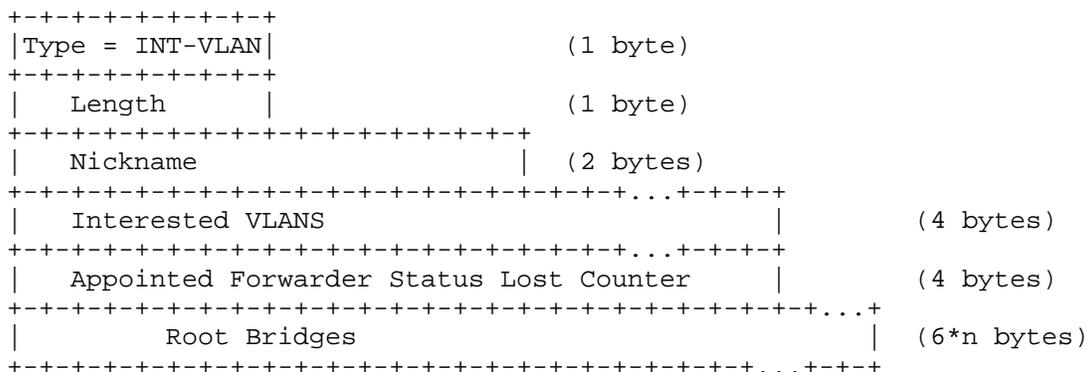
o Type: Router Capability sub-TLV type, set to 10 (INT-VLAN).

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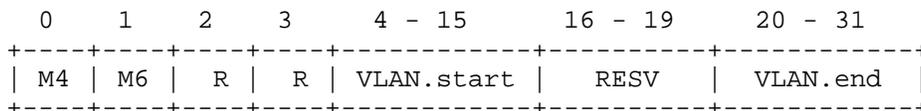
```
        0    1    2    3    4 - 15      16 - 19    20 - 31
      +----+----+----+----+-----------+----------+------------+
      | M4 | M6 | R  | R  | VLAN.start |   RESV   |  VLAN.end  |
      +----+----+----+----+-----------+----------+------------+
```

   - 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 0xFFF, the sub-TVL is interpreted as if
      VLAN.end was 0xFFE. If VLAN.start is less than VLAN.end, the
      sub-TLV is ignored. If both VLAN.start and VLAN.end are 0x000,
      the sub-TLV is ignored. If both VLAN.start and VLAN.end are
      0xFFF then if the nickname field has a valid nickname value

          (0x0001 through 0xFFBF inclusive), that nickname is reserved
          for use in connection with traffic engineered routes to the
          originating RBridge.

     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 announcing 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, which 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 24-bit 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)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...+-+-+-+-+
   |   Interested Labels                              |   (7 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...+-+-+-+-+
   |   Appointed Forwarder Status Lost Counter        |   (4 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...+-+-+-+
   |         Root Bridges                             |   (6*n bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...+-+-+
```
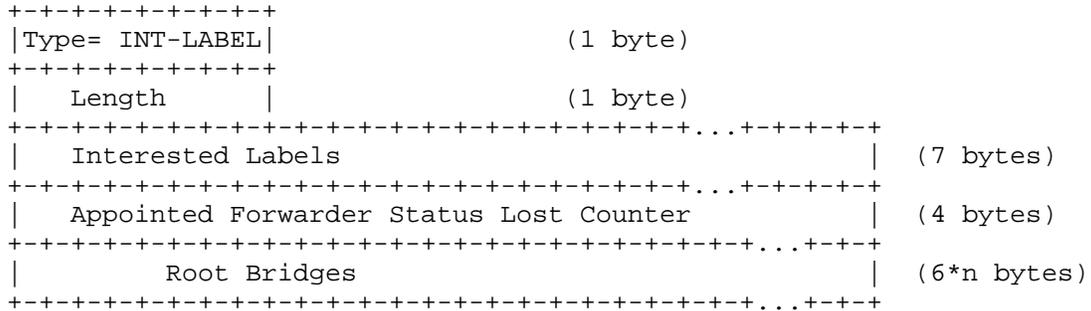
   o Type: Router Capability sub-TLV Type, set to TBD [15 suggested]
      (INT-LABEL).

   o Length: 11 + 6*n where n is the number of root bridge IDs.

   o Interested Labels: The Interested Labels field is seven bytes long
      and formatted as shown below.
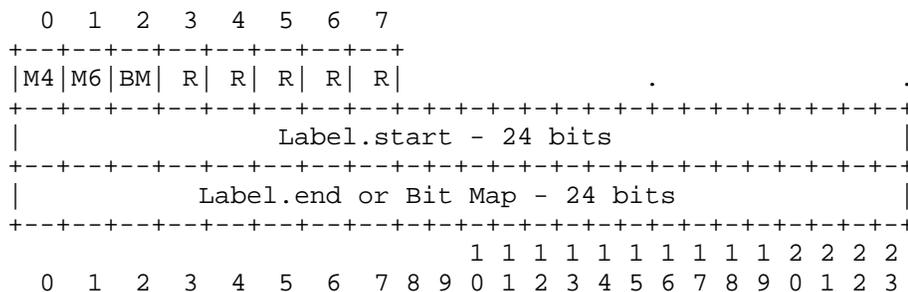
```
   0  1  2  3  4  5  6  7
+--+--+--+--+--+--+--+--+
|M4|M6|BM| R| R| R| R| R|                      .                    .
+--+--+--+--+--+--+--+--+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Label.start - 24 bits                 |
+--+--+--+--+--+--+--+--+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Label.end or Bit Map - 24 bits             |
+--+--+--+--+--+--+--+--+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                               1 1 1 1 1 1 1 1 1 1 2 2 2 2
   0  1  2  3  4  5  6  7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
```

- 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 in 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 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.

2.3.9 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
         +--+--+--...
```

   Note that the bit vector length (BVL) is a seven bit unsigned integer
   field and 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 (see Section 6) in which it occurs. If it

does, it is ignored. If multiple byte-aligned bit vectors are present
in one such sub-TLV, they 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
               0b1000000

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

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)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
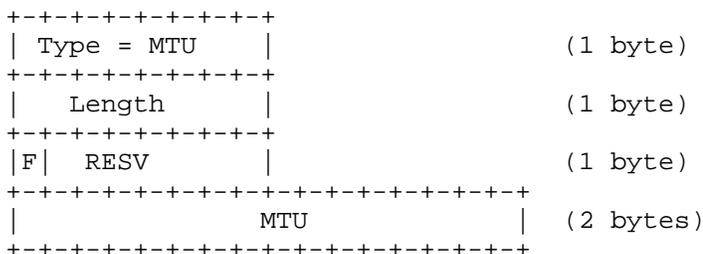
   o Type: Extended Reachability sub-TLV type, set to MTU sub-TLV 28.

   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)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
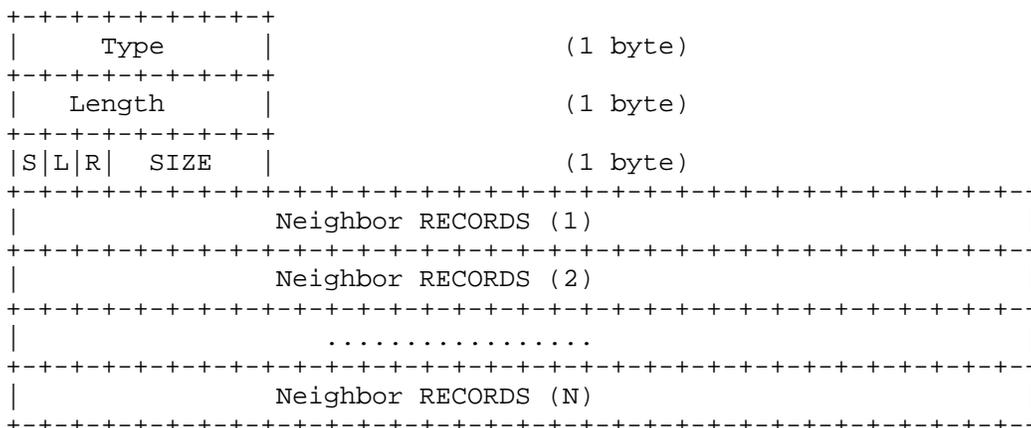
   The information present for each neighbor is as follows:

```
      +-+-+-+-+-+-+-+-+
      |F|   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 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 PDUs are used to optionally determine the MTU on a link
   between ISs as specified in [RFC6325], Section 4.3.2.

   The MTU PDUs have the IS-IS PDU common header (up through the Maximum
   Area Addresses byte) with two new PDU Type numbers as listed in
   Section 6. They also have a 20-byte common fixed MTU PDU header as
   shown below.

```
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |    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 normally 6 bytes but is
   actually set 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 will be
   quite small compared with their minimum length (minimum 1470-byte MTU
   on an 802.3 link, for example), so this will 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 fragment zero PDUs.

4.4 Link State PDUs (LSPs)

   A fragment zero LSP MUST NOT be sent larger than 1470 bytes but a
   larger fragment 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 9 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 "NUM" 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 "NUM" columns are irrelevant to the
      IANA sub-registries.

	Section	sub-TLV#	Grp. Adr.	MT Port	Rtr. Cap.	Ext. Reach	NUM
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	-	-	*
TRILL-PORT-VER	2.2.4	TBD	-	X	-	-	0-1
VLANs-Appointed	2.2.5	TBD	-	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	-	*
MTU	2.4	28	-	-	-	X	0-1
	Section	sub-TLV#	Grp. Adr.	MT Port	Rtr. Cap.	Ext. Reach	NUM

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 TRILL-PORT-VER and TRILL-VER sub-TLVs, which
   are specified herein as MUST be sent as zero and ignored on receipt,
   are allocated based on Standards Action [RFC5226] as modified by

[RFC4020].

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 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 TRILL-PORT-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, support which can be 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 changes, 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 Interested
      VLANs and Spanning Tree Roots sub-TLV including provision for
      reserving a nickname for use in traffic engineering.

   6. Addition of the Interested Labels and Spanning Tree Roots sub-TLV
      to indicate attachment of an IS to a 24-bit fine grained label
      analogous to the existing Interested VLANs and Spanning Tree
      Roots sub-TLV for 12-bit 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 zero.

12. Specify the IANA Considerations policy for reserved and
    capability bits.

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.

   [RFC4020] - Kompella, K. and A. Zinin, "Early IANA Allocation of
        Standards Track Code Points", BCP 100, RFC 4020, February 2005.

   [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.

   [RFCaf] - draft-ietf-trill-rbridge-af, in RFC Editor queue.

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

[ExtendHeader] - draft-ietf-trill-rbridge-extensions, 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.

    [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

Authors' Addresses

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

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

   Ayan Banerjee
   Cisco Systems
   170 West Tasman Drive
   San Jose, CA 95134 USA

   EMail: ayabaner@cisco.com

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

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

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

   Phone: +1-408-333-7149
   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

Copyright, Disclaimer, and Additional IPR Provisions

TRILL Working Group   Donald Eastlake
INTERNET-DRAFT   Mingui Zhang
Intended status: Proposed Standard   Huawei
Updates: 6325, 6327   Anoop Ghanwani
                                                     Brocade
                                              Ayan Banerjee
                                                       Cisco
                                              Vishwas Manral
                                             Hewlett-Packard
Expires: April 30, 2012   October 31, 2011

                  RBridges: Clarifications and Corrections
            <draft-eastlake-trill-rbridge-clear-correct-01.txt>

Abstract

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

   Since the TRILL base protocol was approved in March 2010, active
   development of TRILL has revealed corner cases that could use
   clarification and a few errors in the original RFC 6325. RFCs 6327,
   XXXX, and YYYY, provide clarifications with respect to Adjacency,
   Appointed Forwarders, and the TRILL ESADI protocol. This document
   provide other known clarifications and corrections and updates RFC
   6325 and RFC 6327.

Status of This Memo

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/1id-abstracts.html

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

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links)
standard [RFC6325] provides optimal pair-wise data frame forwarding
without configuration in multi-hop networks with arbitrary topology,
safe forwarding even during periods of temporary loops, and support
for multipathing of both unicast and multicast traffic. TRILL
accomplishes this by using IS-IS (Intermediate System to Intermediate
System) [IS-IS] [RFC1195] [RFC6326] link state routing and
encapsulating traffic using a header that includes a hop count. The
design supports VLANs (Virtual Local Area Networks) and optimization
of the distribution of multi-destination frames based on VLANs and IP
derived multicast groups. Devices that implement TRILL are called
RBridges.

Since the TRILL base protocol [RFC6325] was approved, the active
development of TRILL has revealed corner cases that could use
clarification and a few errors in the original specification document
[RFC6325]. [RFC6327], [RFCXXXX], and [RFCYYYY], provide
clarifications with respect to Adjacency, Appointed Forwarders, and
the TRILL ESADI protocol. This document provide other known
clarifications and corrections and updates [RFC6325].

1.1 Terminology and Acronyms

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

    CFI - Canonical Format Indicator

    DEI - Drop Eligibility Indicator

    OOMF - Overload Originated Multi-destination Frame

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

2. Overloaded and/or Unreachable RBridges

   RBridges may be in overload as indicated by the [IS-IS] overload flag
   in their LSPs. This means that either (1) they are incapable of
   holding the entire link state database and thus do not have a view on
   the entire topology or (2) they have been configured to have the
   overload bit on. Although networks should be engineered to avoid
   actual link state overload, it might occur if a large campus included
   one or more low-end RBridges.

   It is a common operational practice to set the overload bit in an
   [IS-IS] router (such as an RBridge) when performing maintenance on
   that router that might affect its ability for correctly forward
   frames; this will usually leave the router reachable for maintenance
   traffic but transit traffic will not normally be routed through it.
   (Also, in some cases, TRILL provides for setting the overload bit in
   the pseudo node of a link to stop TRILL Data traffic on an access
   link (see Section 4.9.1 of [RFC6325]).)

   [IS-IS] and TRILL make a reasonable effort to do what they can even
   if some RBridge/routers are in overload. They can do reasonable well
   if a few scattered nodes are in overload. However, actual least cost
   paths are no longer assured if any RBridges are in overload.

   An RBridge in overload cannot be trusted to correctly calculate
   distribution trees or correctly perform the Reverse Path Forwarding
   Check. Therefore, it cannot be trusted to forward multi-destination
   TRILL Data frames. It can only appear as a leaf node in a TRILL
   multi-destination distribution tree.

   Frames are not routed through an overloaded RBridge if any other path
   is available, although they may originate or terminate at an
   overloaded RBridge. In addition, TRILL will not route frames over
   links with cost 2**24 - 1; such links are reserved for traffic
   engineered frames the handling of which is beyond the scope of this
   document.

   As a result, a portion of the campus may be unreachable for TRILL
   Data because all paths to it would be through a link with cost 2**24
   - 1. For example, an RBridge RB1 is not reachable by TRILL Data if
   all of its neighbors are connected to RB1 by links with cost 2**24 -
   1. Such RBridges are called "data unreachable".

   The link state database at an RBridge RB1 can also contain
   information on RBridges that are unreachable by IS-IS link state
   flooding due to link or RBridge failures. When such failures
   partition the campus, the RBridges adjacent to the failure and on the
   same side of the failure as RB1 will update their LSPs to show the
   lack of connectivity and RB1 will receive those updates. However,
   LSPs held by RB1 for RBridges on the far side of the failure will not

be updated and may stay around until they time out, which could be
tens of seconds or longer. Since a link is only usable if both ends
declare it up in their LSP, RB1 will be aware of the partition and
will know about the unreachability of nodes on the far side of the
failure. Such nodes are both IS-IS unreachable and data unreachable.

2.1 Distribution Tree Roots

When an RBridge determines what nicknames to use as the roots of the
distribution trees it calculates, it MUST ignore all nicknames held
by RBridges that are in overload or are data unreachable. When
calculating Reverse Path Forwarding checks for multi-destination
frames, an RBridge RB1 can similarly ignore any trees that cannot
reach to RB1 even if other RBridges list those trees as trees those
other RBridges might use. (But see Section 3.)

2.2 Overloaded Receipt of TRILL Data Frames

An RBridge in overload, RB2, will not normally receive unicast TRILL
Data frames unless it is the egress, in which case it processes the
frame normally.

If RB2 receives a unicast TRILL Data frame for which it is not the
egress, perhaps because a neighbor does not yet know it is in
overload, it decrements the Hop Count as usual and discards the frame
if the Hop Count is zero or the egress nickname is illegal. It SHOULD
NOT discard the frame because the egress nickname is unknown as it
might now know about all nicknames due to overload. If any neighbor,
other than the neighbor from which it received the frame, is not
overloaded it MUST attempt to forward the frame to one of those
neighbors.

If RB2 in overload receives a multi-destination TRILL Data frame, RB2
performs the usual Hop Count processing but MUST NOT apply a Reverse
Path Forwarding Check since, due to overload, it might not do so
correctly. RB2 egresses and delivers the frame locally as appropriate
but RB2 MUST NOT forward it (except as an egressed native frame where
RB2 is appointed forwarder).

2.3 Overloaded Origination of TRILL Data Frames

Overloaded origination of unicast frames with know egress and of
multi-destination frames are discussed in the subsections below.

2.3.1 Known Unicast Origination

   When an overloaded RBridge RB2 ingresses or creates a known
   destination unicast TRILL Data frame, it delivers it locally if the
   destination MAC is local. Otherwise RB2 link unicasts it to any
   neighbor RBridge that is not overloaded.

2.3.2 Multi-Destination Origination

   RB2 ingressing or creating a multi-destination TRILL Data frame is
   more complex than for a known unicast frame.

   RB2 would like to hand multi-destination frames it is originating to
   a neighbor RB3 such that that (1) RB3 is not overloaded, (2) RB3 will
   distribute the frame on a tree in which RB2 is a leaf node from RB3,
   and (3) where RB3 will know not to send a copy back to RB2. In the
   absence of criteria 2 and 3, RB2 would received a duplicate copy
   back, which might be confusing. Criteria 2 should not be a problem as
   RBridges in overload can only be leaf nodes for any TRILL
   distribution tree.

   Neighbors of RB2 that meet the criteria and are willing to accept
   such frames advertise this in the TRILL Neighbor TLV in their TRILL
   Hellos by setting a bit associated with the SNPA (MAC address) of RB2
   on the link. (See Section 6.) If no neighbor of RB2 that is not in
   overload offers such service, then RB2 cannot originate multi-
   destination TRILL Data frames, although it can still receive them.

   If RB2 sees this OOMF (Overloaded Origination of Multi-destination
   Frame) service advertised by any of its neighbors on any link to
   which RB2 connects, it selects one such neighbor by a means beyond
   the scope of this document. Assuming RB2 selects RB3 to handle multi-
   destination frames it originates. RB2 MUST advertise in its LSP that
   it might use any of the distribution trees that RB3 advertises it
   might use so that the Reverse Path Forwarding Check will work in the
   rest of the campus.

   RB2 then encapsulates such frames as TRILL data frames to RB3 as
   follows: M bit = 1, Hop Count = 2, ingress nickname = a nickname held
   by RB2, and, since RB2 cannot tell what distribution tree RB3 will
   use, egress nickname = a special nickname indicating an OOMF (see
   Section 6). RB2 then link unicasts this TRILL Data frame to RB3.

   On receipt of such a frame, RB3 does the following:

   - change the egress nickname field to designate a distribution tree
      that RB3 normally uses for which RB2 is a leaf from RB3 (if there
      is no such tree, RB3 MUST NOT offer the OOMF service to RB2),

- change the Hop Count to the value it would normally use if it were
  the ingress, and
- forward the frame on that tree except that it MUST NOT send a copy
  back to RB2.

RB3 MAY rate limit the number of frames for which it is providing
this service by discarding some such frame from RB2. (The provision
of even a limited bandwidth for OOMFs by RB3, for example via a slow
path, may be important to the bootstrapping of services at RB2 or end
stations connected to RB2.)

3. Distribution Tree Updates

When a link state database change causes a change in the distribution
tree(s), there are several possibilities. If a tree root remains a
tree root but the tree changes, then local forwarding and RPFC
entries for that tree should be updated as soon as practical.
Similarly, if a new nickname becomes a tree root, forwarding and RPFC
entries for the new tree should be installed as soon as practical.
However, if a nickname ceases to be a tree root and there is
sufficient room in local tables, it is RECOMMENDED that forwarding
and RPFC entries for the former tree be retained so that any frames
in flight on that tree can still be forwarded.

4. Nickname Selection

   Nickname selection is covered by Section 3.7.3 of [RFC6325]. However,
   the following should be noted:

   1. The second sentence in the second bullet item in Section 3.7.3 of
      [RFC6325] on page 25 is erroneous and is corrected as follows:

      1.a The occurrence of "IS-IS ID (LAN ID)" is replaced with
          "priority".

       1.b The occurrence of "IS-IS System ID" is replaced with "seven
           byte IS-IS ID (LAN ID)".

      The resulting corrected [RFC6325] sentence reads as follows: "If
      RB1 chooses nickname x, and RB1 discovers, through receipt of an
      LSP for RB2 at any later time, that RB2 has also chosen x, then
      the RBridge or pseudonode with the numerically higher priority
      keeps the nickname, or if there is a tie in priority, the RBridge
      with the numerically higher seven byte IS-IS ID (LAN ID) keeps the
      nickname, and the other RBridge MUST select a new nickname."

   2. In examining the link state database for nickname conflicts,
      nicknames held by IS-IS unreachable RBridges MUST be ignored but
      nicknames held by data unreachable RBridges MUST NOT be ignored.

   3. An RBridge may need to select a new nickname, either initially
      because it has none or because of a conflict. When doing so, the
      RBridge MUST consider as available all nicknames that do not
      appear in its link state database or that appear to be held by IS-
      IS unreachable RBridges; however, it SHOULD give preference to
      selecting new nicknames that do not appear to be held by any
      RBridge in the campus, reachable or unreachable, so as to minimize
      conflicts if IS-IS unreachable RBridges later become reachable.

   4. An RBridge, even after it has acquired a nickname for which there
      appears to be no conflicting claimant, MUST continue to monitor
      for conflicts with the nickname or nicknames it holds. It does so
      by checking in LSPs that should update its link state database for
      any of its nicknames held with higher priority by another RBridge
      that is IS-IS reachable. If it finds such a conflict, it MUST
      select a new nickname. (It is possible to receive an LSP that
      should but does not update the link state databse due to
      overflow.)

   5. In the very unlikely case that an RBridge is unable to obtain a
      nickname because all valid nicknames (0x0001 through 0xFFBF
      inclusive) are in use with higher priority by IS-IS reachable
      RBridges, it will be unable to act as an ingress, egress, or tree
      root but will still be able to function as a transit RBridge. Such

an RBridge with no valid nickname MUST announce its nickname in
its LSP as 0x0000. Although it cannot be a tree root, such an
RBridge is included in every distribution tree computed for the
campus. It would not be possible to send an RBridge Channel
message to such an RBridge [Channel].

5. MTU

   Section 5.1 below corrects an Errata in [RFC6325] and Section 5.2
   clarifies the meaning of various MTU numbers.

5.1 MTU PDU Addressing and Processing

   [RFC6325] in Section 4.3.2 incorrectly states that multi-destination
   MTU-probe and MTU-ack TRILL IS-IS PDUs are sent on Ethernet links
   with the All-RBridges multicast address as the Outer.MacDA. As TRILL
   IS-IS PDUs, when multicast on an Ethernet link, they MUST be sent to
   the All-IS-IS-RBridges multicast address.

   As discussed in [RFC6325] and, in more detail, in [RFC6327], MTU-
   probe and MTU-ack PDUs MAY be unicast; however, Section 4.6 of
   [RFC6325] erroneously does not allow for this possibility. It is
   corrected by replacing item numbered "1" in Section 4.6.2 of
   [RFC6325] with the following quoted text to which RBridges MUST
   conform:

   "1. If the Ethertype is L2-IS-IS and the Outer.MacDA is either All-
       IS-IS-RBridges or the unicast MAC address of the receiving
       RBridge port, the frame is handled as described in Section
       4.6.2.1"

   The reference to "Section 4.6.2.1" in the above quoted text is to
   that Section in [RFC6325].

5.2 MTU Values

   MTU values in TRILL key off the originatingLSPBufferSize TLV. In
   layer 3 IS-IS, this defaults to 1492 bytes and is the maximum
   permitted size of LSPs after the eight byte fixed IS-IS PDU header.
   (This header starts with the 0x83 Intradomain Routeing Protocol
   Discriminator byte and ends with the Maximum Area Addresses byte,
   incluside.) Thus the default LSP size including this header is 1500
   bytes. In TRILL, originatingLSPBufferSize defaults to 1470 bytes,
   allowing 22 bytes of additional headroom to accomodate legacy device
   with, for example, the classic Ethernet maximum MTU.

   More TBD - talk about exact meaning of MTU values for various link
   technologies

6. The CFI / DEI Bit

   In May 2011, the IEEE promulgated [802.1Q-2011] which changes the
   meaning of the bit between the priority and VLAN ID bits in the
   payload of C-VLAN tags. Previously this bit was called the CFI
   (Canonical Format Indicator) bit and had a special meaning in
   connection with IEEE 802.5 (Token Ring) frames. Now, under
   [802.1Q-2011], it is a DEI (Drop Eligibility Indicator) bit, similar
   to that bit in S-VLAN (and B-VLAN) tags where this bit has always
   been a DEI bit.

   The TRILL base protocol specification [RFC6325] assumed, in effect,
   that the link by which end stations are connected to RBridges and the
   virtual link provided by the TRILL Data frame encapsulation are IEEE
   802.3 Ethernet links on which the CFI bit is always zero. Should an
   end station be attached by some other type of link, such as an FDDI
   (Fiber Distributed Data Interface) or Token Ring link, [RFC6325]
   implicitly assumed that such frames would be canonicalized to 802.3
   frames before being ingressed and similarly, on egress, such frames
   would be converted from 802.3 to the appropriate frame type for the
   link. Thus, [RFC6325] required that the CFI bit in the Inner.VLAN
   always be zero.

   However, for RBridges with ports conforming to the change
   incorporated in the IEEE 802.1Q-2011 standard, the bit in the
   Inner.VLAN, now a DEI bit, MUST be set to the DEI value provided by
   the EISS interface on ingressing a native frame. Similarly, this bit
   MUST be provided to the EISS when transiting or egressing a TRILL
   Data frame. The exact effect on the C-VLAN Outer.VLAN DEI and
   priority bits and whether or not an Outer.VLAN appears at all on the
   wire for output frames depends on output port configuration.

   RBridge campuses with a mixture of ports, some compliant with
   [802.1Q-2011] and some compliant with pre-802.1Q-2011, especially if
   they have actual Token Ring links, may operate incorrectly and may
   corrupt data, just as a bridged LAN with such mixed ports would.

7. Graceful Restart

   RBridge SHOULD support the features specified in [RFC5306] which
   describes a mechanism for a restarting IS-IS router to signal to its
   neighbors that it is restarting, allowing them to reestablish their
   adjacencies without cycling through the down state, while still
   correctly initiating link state database synchronization.

8. Update to RFC 6327

   [RFC6327] provides for multiple states of the potential adjacency
   between two RBridges. It makes clear that an adjacency is reported in
   LSPs in the "Report" state. LSP transmission and synchronization,
   however, performed in both the "Two-Way" and "Report" states.

8. IANA Considerations

   IANA is requested to allocate the previously reserved nickname 0xXXXX
   for use in the TRILL Header egress nickname field to indicate an
   Overload Originated Multi-destination Frame (OOMF).

   IANA is requested to allocate bit TBD (bit 1, the bit adjacent to the
   F bit suggested) from the seven currently reserved (RESV) bits in the
   per neighbor "Neighbor RECORD" in the TRILL Neighbor TLV [RFC6326].
   This bit indicates that the RBridge sending the TRILL Hello will
   provide the OOMF forwarding service described in Section 2.3 to such
   frames originated by the RBridge whose SNPA (MAC address) appears in
   that Neighbor RECORD.

9. Security Considerations

   This memo improves the documentation of the TRILL standard and
   corrects some errors in [RFC6325]. It does not change the security
   considerations of the TRILL base protocol for which see Section 6 of
   [RFC6325].

Normative References

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

    [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.

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

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

    [RFC5306] - Shand, M. and L. Ginsberg, "Restart Signaling for IS-IS",
        RFC 5306, October 2008.

    [RFC6325] - Perlman, R., Eastlake 3rd, 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.

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

Informative References

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

    [RFCXXXX] - R. Perlman, D. Eastlake, Y. Li, A. Banerjee, H. Fangwei,
        "RBridges: Appointed Forwarders", draft-ietf-trill-rbridge-af,
        in the RFC Editor's queue.

    [RFCYYYY] - H. Zhai, F. Hu, R. Perlman, D. Eastlake, "RBridges: The
        ESADI Protocol", draft-hu-trill-rbridge-esadi, work in
        progress.

Authors' Addresses

   Donald Eastlake
   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

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

   Phone: +1-408-333-7149
   EMail: anoop@alumni.duke.edu

   Ayan Banerjee
   Cisco Systems
   170 West Tasman Drive
   San Jose, CA 95134 USA

   EMail: ayabaner@cisco.com

   Vishwas Manral
   HP Networking
   19111 Pruneridge Avenue
   Cupertino, CA 95014 USA

   Tel: +1-408-477-0000
   EMail: vishwas.manral@hp.com

Appendix: Change Record

    This appendix summarizes changes between versions of this draft.

    RFC Editor: Please delete this Appendix before publication.

From -00 to -01

    1. Add Section updating [RFC6327].

    2. Add some material to Section 5.2 on MTUs.

    3. Minor editorial changes.

Copyright and IPR Provisions

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

Abstract

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

Status of this Memo

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

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

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

   This Internet-Draft will expire on January 5, 2012.

Copyright Notice

   Copyright (c) 2011 IETF Trust and the persons identified as the

Table of Contents

1. Problem Statement

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

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

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

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

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

2. Pseudonode Nickname

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

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

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

3. LSP Announcement

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

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

4. Unicast TRILL Data Frames Processing

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

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

4.1. Ingress processing

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

4.2. Egress processing

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

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

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

4.2.1. Unicasting to VLAN-x Forwarder

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

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

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

4.2.2. Multicasting to VLAN-x forwarder

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

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

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

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

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

4.2.3. Comparison

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

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

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

5. TLV Extensions for Pseudonode Nickname

5.1. Pseudonode Nickname Capability in Hellos

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

Pseudonode Nickname capability TLV

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

Figure 1

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

Other bits and fields refer to [TrillISIS].

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

5.2. Pseudonode Nickname TLV

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

Pseudonode Nickname TLV

```
   +-+-+-+-+-+-+-+-+
   |Type= PSEU-NICK|                         (1 byte)
   +-+-+-+-+-+-+-+-+
   |    Length     |                         (1 byte)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 PSEUDONODE NICKNAME RECORDS (1)              |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      ..................                      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 PSEUDONODE NICKNAME RECORDS (n)              |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
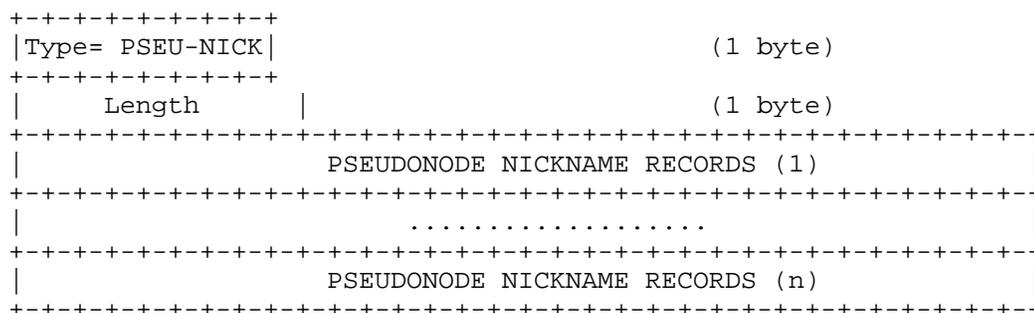
   where each pseudonode nickname record is of the form:

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


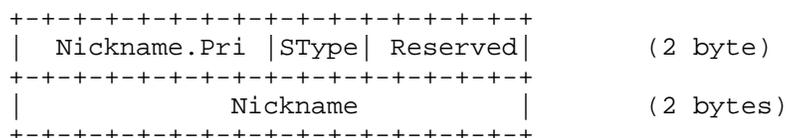   Figure 2

   o Type: Pseudonode Nickname Type, TBD (NICKNAME).

   o Length: 4*N, where N is the number of pseudonode nickname records
   present.

   o SType: An 3-bit unsigned integer sub-type for nickname. If this
   nickname is pseudonode nickname, value of this field is 1.

   o Nickname.Pri: An 8-bit unsigned integer priority to hold a nickname
   as specified in Section 3.7.3 of [RFCtrill].

   o Nickname: This is an unsigned 16-bit integer as specified in
   Section 3.7 of [RFCtrill].

5.2.1. Pseudonode Nickname TLV in Hellos

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

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

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

5.2.2. Pseudonode Nickname TLV in DRB's LSPs

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

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

6. Security Considerations

7. Acknowledgements

8. References

8.1. Normative references

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

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

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

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

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

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

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

8.2. Informative References

Authors' Addresses

    Hongjun Zhai
    ZTE Corporation
    68 Zijinghua Road
    Nanjing 200012
    China

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

    Fangwei Hu
    ZTE Corporation
    889 Bibo Road
    Shanghai 201203
    China

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

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

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

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

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

TRILL Working Group   Donald Eastlake
INTERNET-DRAFT   Huawei
Intended status: Proposed Standard   Anoop Ghanwani
                                                               Brocade
                                                         Vishwas Manral
                                                          HP Networking
                                                        Caitlin Bestler
                                                                Quantum
Expires: April 23, 2012   October 24, 2011

                    RBridges: TRILL Header Extensions
                 <draft-ietf-trill-rbridge-extension-00.txt>

Abstract

   The TRILL base protocol standard specifies minimal hooks to safely
   support TRILL Header extensions. This document specifies an initial
   extension providing additional flag bits and specifies one of those
   bits.

Status of This Memo

Table of Contents

1. Introduction

   The base TRILL protocol standard [RFC6325] provides a TRILL Header
   extension feature, called "options" in Section 3.8 of [RFC6325], and
   describes minimal hooks to safely support header extension. But,
   except for the first two bits, it does not specify the structure of
   the extension to the TRILL Header nor the details of any particular
   extension. This document specifies an initial extension providing
   additional flag bits and specifies one of those bits. Additional
   extensions, including TLV (Type, Length, Value) encoded options, may
   be specified in later documents.

   Section 2 below describes some general principles of TRILL Header
   Extensions and an initial extension. Section 3 describes a specific
   flag in this initial extension.

1.1 Conventions used in this document

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

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

2. TRILL Header Extensions

   The base TRILL Protocol includes a feature for extension of the TRILL
   Header (see [RFC6325] Sections 3.5 and 3.8). The 5-bit Op-Length
   header field gives the length of the extension to the TRILL Header in
   units of 4 octets, which allows up to 124 octets of header extension.
   If Op-Length is zero there is no header extension present; else, this
   area follows immediately after the Ingress Rbridge Nickname field of
   the TRILL Header. The first 32-bit word of the optional extensions
   area consists of an extended flags area specified in this document.

   As described below, provision is made for hop-by-hop flags, which
   might affect any RBridge that receives a TRILL Data frame with such a
   flag set, ingress-to-egress flags, which would only necessarily
   affect the RBridge(s) where a TRILL frame is decapsulated, and a
   third type of intermediate flag affecting an as yet unspecified class
   of RBridges, for example border RBridges in a TRILL campus extended
   to support multi-level IS-IS. Provision is also made for both
   "critical" and "non-critical" flags.

   Any RBridge receiving a frame with a critical hop-by-hop extension
   that it does not implement MUST discard the frame because it is
   unsafe to process the frame without understanding such a critical
   extension. Any egress RBridge receiving a frame with a critical
   ingress-to-egress or hop-by-hop extension it does not implement MUST
   drop the frame if it is a known unicast frame; if it is a multi-
   destination TRILL Data frame, then it MUST NOT be egressed at that
   RBridge but it is still forwarded on the distribution tree. Non-
   critical extensions can be safely ignored.

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

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

2.1 RBridge Extended Flag Handling Requirements

   All RBridges MUST check whether there are any critical flags set that
   are necessarily applicable to their processing of the frame. To
   assist in this task, critical summary bits are provided that cover
   not only the extended flags specified herein but will cover any
   further extensions specified in future documents [Options]. If an
   RBridge does not implement all critical flags in a TRILL Data frame,
   it MUST discard the frame or, in some circumstances as described
   above for certain multi-destination frames, continue to forward the
   frame but MUST NOT egress the frame.

   In addition, a transit RBridge:

   o MAY set or clear hop-by-hop flags as specified for such flags;
   o MUST adjust the length of the extensions area, including changing
      Op-Length in the TRILL header, as appropriate if it adds or
      removes the word of extended header flags;
   o MUST, if it adds the word of extended header flags or changes any
      critical flags, correctly set the critical summary bits in the
      extended header flags word;
   o MUST NOT remove the extended header flags word unless it is all
      zero (either on arrival or after permitted modifications);
   o MUST NOT set or clear ingress-to-egress or reserved extended
      header flags except as specifically permitted in the specification
      of the flag.

2.2 No Critical Surprises

   RBridges advertise the extended header flags they support in IS-IS
   PDUs. Unless an RBridge advertises support for a critical extended
   header flag, it will not normally receive frames with that flag set.
   An RBridge is not required to support any extensions.

   An RBridge SHOULD NOT set a critical extended flag in a frame unless,

   - for a critical hop-by-hop extended header flag, it has determined
      that the next hop RBridge or RBridges that will accept the frame
      support that flag,
   - for a critical ingress-to-egress extended header flag, it has
      determined that the RBridge or RBridges that will egress the frame
      support that flag, or
   - for a critical reserved extended header flag, it may set such a
      flag only if it understands which RBridges it is applicable to and
      has determined that those RBridges that will accept the frame
      support that flag.

   "SHOULD NOT" is specified since there may be cases where it is

    acceptable for those frames, particularly for the multi-destination
    case, to be discarded by any RBridges that do not implement the
    extended flag.

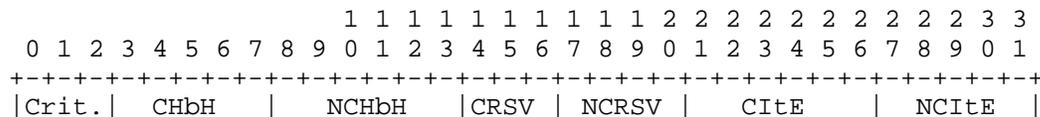2.3 Extension Header Flags

    If any extensions are present in a TRILL Header, as indicated by a
    non-zero Op-Length field, the first 32 bits of the extensions area
    consist of extended header flags, as described below. The remainder
    of the extensions area, if any, after this initial 32 bits, will be
    specified in later documents [Options].

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

    Bit(s)   Description

      0-3 Crit.: Critical summary bits.
          0 CHbHS: Critical Hop-by-Hop extension(s) are present.
          1 CItES: Critical Ingress-to-Egress extension(s) are present.
          2 CRSVS: Critical reserved extension(s) are present.

     3-7   CHbH: Critical Hob-by-Hop extended Flag bits.
     8-13 NCHbH: Non-critical Hop-by-Hop extended Flag bits.

     14-16 CRSV: Critical Reserved extended Flag bits.
     17-20 NCRSV: Non-critical Reserved extended Flag bits.

     21-26 CItE: Critical Ingress-to-Egress extended Flag bits.
     27-31 NCItE: Non-critical Ingress-to-Egress extended Flag bits.

    These are illustrated below:

                        1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+
     |Crit.| CHbH | NCHbH |CRSV | NCRSV | CItE | NCItE |

    For TRILL Data frames with extensions present, any transit RBridge
    MUST transparently copy the extended flags word, except as permitted
    by an extension implemented by that RBridge.

2.3.1 Critical Summary Bits

   The top three bits of the extensions area, bits 0, 1, and 2 above,
   are called the critical summary bits. They summarize the presence of
   critical extensions as follows:

   CHbHS: If the CHbHS (Critical Hop by Hop Summary) bit is one, one or
      more critical hop-by-hop extensions are present. These might be
      critical hop-by-hop extended header flags or critical hop-by-hop
      extensions after the first word in the extensions area. Transit
      RBridges that do not support all of the critical hop-by-hop
      extensions present, for example an RBridge that supported no hop-
      by-hop extensions, MUST drop the frame. If the CHbH bit is zero,
      the frame is safe, from the point of view of extensions
      processing, for a transit RBridge to forward, regardless of what
      extensions that RBridge does or does not support.

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

   CRSVS: If the CRSVS (Critical Reserved Summary) bit is a one, one or
      more critical extensions are present that are reserved to apply to
      a class of RBridges to be specified in the future, for example
      border RBridges in a TRILL campus extended to support multi-level
      IS-IS. This class will be a subset of transit RBridges. RBridges
      in this class MUST drop frames with the CRSVS bit set unless they
      implement all critical hop-by-hop and all critical reserved
      extensions present in the frame.

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

2.3.2 Extended Header Flags

   CHbH, bits 3 to 7, are Critical Hob-by-Hop extended Flag bits.

   NCHbH bits 8 to 13, are Non-critical Hop-by-Hop extended Flag bits.

   CRSV, bits 14 to 16, are Critical Reserved extended Flag bits.

   NCRSV, bits 17 to 20, are Non-critical Reserved extended Flag bits.

   CItE, bits 21 to 26, are Critical Ingress-to-Egress extended Flag
   bits.

   NCItE, bits 27 to 31, are Non-critical Ingress-to-Egress extended
   Flag bits.

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

2.4 Conflict of Extensions

   It would be possible for TRILL Header extension flags to conflict.
   Two or more extension flags could be present in a frame that direct
   an RBridge processing the frame to do conflicting things or to change
   its interpretation of later parts of the frame in conflicting ways.
   Such conflicts are resolved by applying the following rules in the
   order given and stopping with the first one that applies:

   1. Any frame containing extensions that require mutually incompatible
      changes in way later parts of the frame, after the extensions
      area, are interpreted or structured MUST be discarded. (Such
      extensions will be critical extensions, normally hop-by-hop
      critical extensions.)

   2. Critical extensions override non-critical extensions.

   3. Within each of the two categories of critical and non-critical
      extensions, the extension appearing first in lexical order in the
      frame always overrides an extension appearing later in the frame.
      Extended flags with lower bit numbers are considered to have
      occurred before extended flags with higher bit numbers. Thus a
      conflict between a hop-by-hop extended flag and an ingress-to-
      egress extended flag is resolved in favor of the hop-by-hop
      extended flag.

3. Specific Extended Header Flag

   The table below shows the state of TRILL Header extended flag
   assignments. See Section 6 for IANA Considerations.

```
      Bits    Purpose                    Section
      ------------------------------------------
       0-2    Critical Summary Bits       2.3.1
       3-7    available for critical hop-by-hop flags
        8     RBridge Channel Alert Flag  3.1
       9-13   available for non-critical hop-by-hop flags
      14-16   available for critical reserved flags
      17-20   available for non-critical reserved flags
      21-26   available for critical ingress-to-egress flags
      27-31   available for non-critical ingress-to-egress flags
```

              Table 1. Extended Header Flags Area

3.1 The RBridge Channel Alert Extended Flag

   The RBridge Channel Alert Extended Flag indicates that the frame is
   an RBridge Channel frame [Channel] that requests processing at each
   hop. This is a non-critical hop-by-hop flag. It is intended to alert
   transit RBridges that implement this extension and to assist in the
   implementation of features such as a record route message.

4. Additions to IS-IS

RBridges use IS-IS LSP PDUs to inform other RBridges which extensions they support. The IS-IS PDU(s), TLV(s), or sub-TLV(s) used to encode and advertise this information are specified in a separate document [RFC6326bis].

5. IANA Considerations

IANA is requested to create a subregistry within the TRILL Parameters registry: The "TRILL Extended Header Flags" subregistry, that is initially populated as specified in Table 1 in Section 3. References in that table to sections of this document are to be replaced in the IANA subregistry by references to this document as an RFC.

New TRILL Extended Header Flags are allocated by Standards Action [RFC5226] as modified by [RFC4020].

6. Security Considerations

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

For security considerations related to extended header flags, see the document where the flag is specified.

It is important that the critical summary bits in the extended header flags word be set properly. If set when critical extensions of the appropriate category are not present, frames may be unnecessarily discarded. If not set when critical extensions are present, frames may be mishandled or corrupted and intended security policies, such as VLAN separation, may be violated.

The RBridge Channel Alert extended flag specified herein has no special security considerations. Implementations should keep in mind that it might be erroneously set in a frame. If found set in a frame that is not an RBridge Channel message [Channel], this flag MAY be cleared and should have no effect except, possibly, delaying processing of the frame. If erroneously omitted from a frame, desired per hop processing for the frame may not occur.

7. Acknowledgements

The following are thanked for their contributions: Thomas Narten.

8. Normative References

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

   [RFC4020] - Kompella, K. and A. Zinin, "Early IANA Allocation of
        Standards Track Code Points", BCP 100, RFC 4020, February 2005.

   [RFC5226] - Narten, T. and H. Alvestrand, "Guidelines for Writing an
        IANA Considerations Section in RFCs", BCP 26, RFC 5226, May
        2008.

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

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

   [RFC6326bis] - draft-eastlake-isis-rfc6326bis, work in progress.

9. Informative References

   [Options] - draft-ietf-trill-rbridge-options, work in progress.
        - draft-eastlake-trill-rbridge-more-options, work in progress.

Change History

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

   Version 00 of this draft is an extract and simplification of draft-
   ietf-trill-rbridge-options-05.txt as discussed at the TRILL WG
   meeting at IETF 81 and on the TRILL WG mailing list.

Authors' Addresses

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

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

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

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

    Vishwas Manral
    HP Networking
    19111 Pruneridge Avenue
    Cupertino, CA 95014 USA

    Tel: +1-408-477-0000
    EMail: vishwas.manral@hp.com

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

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

Copyright and IPR Provisions

Network Working Group   M. Wasserman
Internet-Draft   Painless Security
Intended status: Standards Track   D. Eastlake
Expires: April 26, 2012   D. Zhang
                                                   Huawei Technologies
                                                     October 24, 2011

         Transparent Interconnection of Lots of Links (TRILL) over IP
                     draft-mrw-trill-over-ip-00.txt

Abstract

   The Transparent Interconnection of Lots of Links (TRILL) protocol is
   implemented by devices called Routing Bridges (RBridges). TRILL
   supports both point-to-point and multi-access links and is designed
   so that a variety of link protocols can be used between RBridge
   ports. This document standardizes a methods for encapsulating TRILL
   in UDP/IP(v4 or v6).

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 26, 2012.

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. Requirements Terminology

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

2. Introduction

   RBridges are devices that implement the IETF TRILL protocol [RFC6325]
   [RFC6326] [RFC6327].

   RBridges provide transparent forwarding of frames within an arbitrary
   network topology, using least cost paths for unicast traffic. They
   support VLANs and multipathing of unicast and multi-destination
   traffic. They use IS-IS link state routing and a hop count. The are
   compatible with IEEE customer bridges, and can incrementally replace
   them.

   Two or more RBridges can communicate over a variety of different link
   types, such as Ethernet [RFC6325] or PPP [RFC6361].

   This document defines a method for RBridges to communicate over UPD/
   IP(v4 or v6). TRILL over IP will allow remote, Internet-connected
   RBridges to form a single RBridge campus, or multiple TRILL over IP
   networks within a campus to be connected via a TRILL over IP
   backbone.

   TRILL over IP connects RBridge ports using IPv4 or IPv6 as a
   transport in such a way that the ports appear to TRILL to be
   connected by a single link. The link will be a multi-access link if
   more than two RBridge ports are connected to a single TRILL over IP
   link.

   To support cases where RBridges are connected via links (such as the
   public Internet) that are not under the same administrative control
   as the TRILL campus, this document specifies the use of Datagram
   Transport Layer Security (DTLS) [RFC4327] to secure communication
   between RBridges running TRILL over IP.

3. Use Cases for TRILL over IP

   In this document, we consider two use cases that are typical of
   situations where network administrators may choose to use TRILL over
   an IP network: a remote office scenario, and an IP backbone scenario.

3.1. Remote Office Scenario

   In the Remote Office Scenario, a remote TRILL network is connected to
   a TRILL campus across a multihop non-TRILL IP network, such as the
   public Internet. The TRILL network in the remote office becomes a
   logical part of TRILL campus, and nodes in the remote office can be
   attached to the same VLANs as local campus nodes. In many cases, a
   remote office may be attached to the TRILL campus by a single pair of
   RBridges, one on the campus end, and the other in the remote office.
   In this use case, the TRILL over IP link will often cross logical and
   physical IP networks that do not support TRILL, and are not under the
   same administrative control as the TRILL campus.

3.2. IP Backbone Scenario

   In the IP Backbone Scenario, TRILL over IP is used to connect a
   number of TRILL networks within a single TRILL campus. For example,
   a TRILL over IP backbone could be used to connect multiple TRILL
   networks on different floors of a large building, or to connect TRILL
   networks in separate buildings of a multi-building site. In this use
   case, there may often be several TRILL RBridges on a single TRILL
   over IP link, and the the IP link(s) used by TRILL over IP are
   typically under the same administrative control as the rest of the
   TRILL campus.

3.3. Important Properties of the Scenarios

   There are a number of differences between the two scenarios listed
   above, some of which drive features of this specification. These
   differences are especially pertinent to the security requirements of
   the solution, how multicast data frames are handled, and how the
   RBridges discover each other.

3.3.1. Security Requirements

   In the IP Backbone Scenario, TRILL over IP is used between a number
   of RBridges, on a network link that is in the same administrative
   control as the remainder of the TRILL campus. While it is desirable
   in this scenario to prevent the association of rogue RBridges, this
   can be accomplished using existing IS-IS security mechanisms. There
   may be no need to protect the data traffic, beyond any protections
   that are already in place on the local network.

   In the Remote Office Scenario, TRILL over IP may run over a network
   that is not under the same administrative control as the TRILL
   network. Nodes on the network may think that they are sending
   traffic locally, while that traffic is actually being sent, in a
   UDP/IP tunnel, over the public Internet. It is necessary in this

scenario to protect user privacy, as well as ensuring that no
unauthorized RBridges can gain access to the RBridge campus. The
data privacy requirement is addressed by the use of DTLS for both
IS-IS frames and data frames between RBridges in this scenario.

3.3.2. Multicast Handling

In the IP Backbone scenario, native mutlicast may be supported on the
TRILL over IP link. If so, it will be used to send TRILL IS-IS and
multicast data frames, as discussed later in this document.

In the Remote Office Scenario, there will often be only one pair of
RBridges connecting a given site, and even when multiple RBridges are
used to connect a Remote Office to the TRILL campus, the intervening
network may not provide reliable (or any) multicast connectivity.
Also, there is no suitable way to provide data privacy for multicast
traffic. For all of these reasons, the connections between local and
remote RBridges will be treated like point-to-point links, and all
TRILL IS-IS control messages and multicast data frames that are
transmitted between the Remote Office and the TRILL campus will be
serialized, as discussed later in this document.

3.3.3. RBridge Discovery

In the IP Backbone Scenario, RBridges that use TRILL over IP will use
the normal TRILL IS-IS Hello mechanisms to discover the existence of
other RBridges on the link, and to establish authenticated
communication with those RBridges.

In the Remote Office Scenario, a DTLS session will need to be
established between RBridges before TRILL IS-IS traffic can be
exchanged, as discussed below. In this case, one of the RBRidges
will need to be configured to establish a DTLS session with the other
RBridge. This will typically be accomplished by configuring the
RBridge at a Remote Office to initiate a DTLS session, and subsequent
TRILL exchanges, with an TRILL over IP-enabled RBridge attached to
the TRILL campus.

4. TRILL Frame Formats

To support the TRILL base protocol standard [RFC6325]. , two types of
frames will be transmitted between RBridges: TRILL Data frames and
TRILL IS-IS frames.

4.1. TRILL Data Frame

   The on-the-wire form of a TRILL Data frame in transit between two
   neighboring RBridges is as shown below:

      +--------------+----------+----------------+-----------+
      | TRILL Data | TRILL | Encapsulated | Link |
      | Link Header | Header | Native Frame | Trailer |
      +--------------+----------+----------------+-----------+
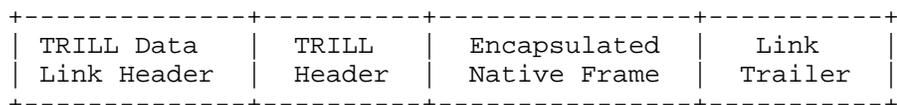
   Where the Encapsulated Native Frame is in Ethernet frame format with
   a VLAN tag but with no trailing Frame Check Sequence (FCS).

4.2. TRILL IS-IS Frame

   TRILL IS-IS frames are formatted on-the-wire as follows:

      +--------------+---------------+-----------+
      | TRILL IS-IS | TRILL IS-IS | Link |
      | Link Header | Payload | Trailer |
      +--------------+---------------+-----------+
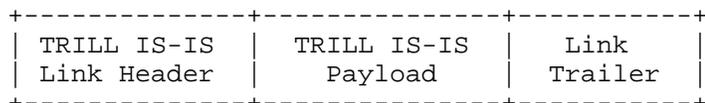
   The Link Header and Link Trailer in these formats depend on the
   specific link technology. The Link Header usually contains one or
   more fields that distinguish TRILL Data from TRILL IS-IS. For
   example, over Ethernet, the TRILL Data Link Header ends with the
   TRILL Ethertype while the TRILL IS-IS Link Header ends with the L2-
   IS-IS Ethertype; on the other hand, over PPP, there are no Ethertypes
   but PPP protocol code points are included that distinguish TRILL Data
   from TRILL IS-IS.

   In TRILL over IP, we will use UDP/IP (v4 or v6) as the link header,
   and the TRILL frame type will be determined based on the UDP port
   number. In TRILL over IP, no Link Trailer is specified, although one
   may be added when TRILL over IP packets are encapsulated for
   transmission on a network (e.g. Ethernet).

5. Link Protocol Specifics

   TRILL Data packets can be unicast to a specific RBridge or multicast
   to all RBridges on the link. TRILL IS-IS packets are always
   multicast to all other RBridge on the link. On Ethernet links, the
   Ethernet multicast address All-RBridges is used for TRILL Data and

   All-IS-IS-RBridges for TRILL IS-IS.

   To properly handle TRILL base protocol frames on a TRILL over IP
   link, either native multicast mode must be enabled on that link, or
   multicast must be simulated using serial unicast, as discussed below.

   In TRILL Hello PDUs used on TRILL IP links, the IP addresses of the
   connected IP ports are their SNPA addresses. Thus, all TRILL
   Neighbor TLVs in such Hellos MUST specify that the size of the SNPA
   is 4-bytes for an IPv4 link or 16-bytes for an IPv6 link
   [rfc6326bis]. Note that SNPA addresses and their size are
   independent of TRILL System IDs which are 6-bytes.

6. Port Configuration

   Each RBridge port that is to be used for a TRILL over IP link MUST
   have at least one IP (v4 or v6) address. Implementations MAY allow a
   single physical port to operate as multiple IPv4 and/or IPv6 logical
   ports.

   TBD: MUST be able to configure list of IP addresses for serial
   unicast. MUST be able to configure non-standard IP multi-cast
   addresses.

7. TRILL over UDP/IP Format

   The general format of a TRILL over UDP/IP packet is shown below.

       +----------+--------+----------------------+
       | IP | UDP | TRILL |
       | Header | Header | Payload |
       +----------+--------+----------------------+
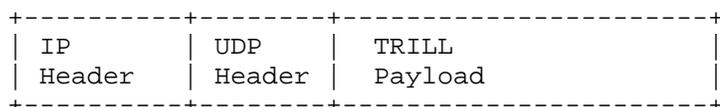
8. Handling Multicast

8.1. Multicast of TRILL IS-IS Packets

8.2. Multicast Data Frames

9. Use of DTLS

   All RBridges that support TRILL over IP MUST implement DTLS and

support the use of DTLS to secure both TRILL IS-IS and data traffic.
When DTLS is used to secure a TRILL over IP link, the DTLS session
MUST be fully established before any TRILL IS-IS or data frames are
exchanged.

RBridges that implement TRILL over IP MUST support the use of
certificates for DTLS authentication, and MUST support the following
algorithm:

o TLS_RSA_WITH_AES_128_CBC_SHA [RFC5246]

RBridges that support TRILL over IP MAY support the use of pre-shared
keys for DTLS authentication. If pre-shared keys are supported, the
following cryptographic algorithms MUST be supported for use with
pre-shared keys:

o TLS_PSK_WITH_AES_128_CBC_SHA [RFC5246]

10. MTU Considerations

   TBD

11. Middlebox Considerations

   TBD

12. Security Considerations

   TRILL over IP is subject to all of the security considerations for
   the base TRILL protocol. In addition, there are specific security
   requirements for different TRILL deployment scenarios, as discussed
   in the "Use Cases for TRILL over IP" section above.

   This document specifies that all RBridges that support TRILL over IP
   MUST implement DTLS, and makes it clear that it is both wise and good
   to use DTLS in all cases where a TRILL over IP link will traverse a
   network that is not under the same administrative control as the rest
   of the TRILL campus. DTLS is necessary, in these cases to protect
   the privacy and integrity of data traffic.

   TRILL over IP is completely compatible with the use of IS-IS
   security, which can be used to authenticte RBridges before allowing
   them to join a TRILL campus. This is sufficient to protect against
   rogue RBridges, but is not sufficient to protect data frames that may
   be sent, in UDP/IP tunnels, outside of the local network, or even

across the public Internet. To protect the privacy and integrity of
that traffic, use DTLS.

In cases were DTLS is used, the use of IS-IS security may not be
necessary, but there is nothing about this specification that would
prevent using both DTLS and IS-IS security together. In cases where
both types of security are enabled, implementations MAY allow users
to configure a single shared key that will be used for both
mechanisms.

13. IANA Considerations

IANA has allocated the following UDP Ports for the TRILL IS-IS and
Data channels:

```
        UDP Port          Protocol

        (TBD)             TRILL IS-IS Channel
        (TBD)             TRILL Data Channel
```

IANA has allocated one IPv4 and one IPv6 multicast address, as shown
below, which correspond to the All-RBridges multicast MAC addresses
that the IEEE Registration Authority has assigned for TRILL.

[Values recommended to IANA:]

```
        TRILL name        IPv4             IPv6

        All-RBridges      233.252.14.0     FF0X:0:0:0:0:0:0:205
```

Note: when these IPv4 and IPv6 multicast addresses are used and the
resulting IP frame is sent over Ethernet, the usual IP derived MAC
address is used.

[Need to discuss scopes for IPv6 multicast (the "X" in the addresses)
somewhere. Default to "site" scope but MUST be configurable?]

14. Acknowledgements

This document was written using the xml2rfc tool described in RFC
2629 [RFC2629].

The following people have provided useful feedback on the contents of this document: Sam Hartman.

15. References

15.1. Normative References

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

   [RFC4327] Dubuc, M., Nadeau, T., Lang, J., and E. McGinnis, "Link
              Management Protocol (LMP) Management Information Base
              (MIB)", RFC 4327, January 2006.

   [RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security
              (TLS) Protocol Version 1.2", RFC 5246, August 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.

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

   [RFC6361] Carlson, J. and D. Eastlake, "PPP Transparent
              Interconnection of Lots of Links (TRILL) Protocol Control
              Protocol", RFC 6361, August 2011.

15.2. Informative References

   [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629,
              June 1999.

Authors' Addresses

Margaret Wasserman
Painless Security
356 Abbott Street
North Andover, MA   01845
USA

Phone: +1 781 405-7464
Email: mrw@painless-security.com
URI: http://www.painless-security.com

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

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

Dacheng Zhang
Huawei Technologies
Q14, Huawei Campus
No.156 Beiqing Rd.
Beijing, Hai-Dian District   100095
P.R. China

Phone:
Email: zhangdacheng@huawei.com
URI:

TRILL Working Group   Tissa Senevirathne
Internet Draft   Dinesh G Dutt
Intended status: Standards Track   CISCO
                                                Vishwas Manral
                                                  HP Networking

                                              October 20, 2011

Expires: April 2012

                    ICMP based OAM Solution for TRILL
                       draft-tissa-trill-oam-00.txt

Status of this Memo

Copyright Notice

Abstract

   This document presents a solution suite for TRILL data plane
   monitoring and failure detection. Methods presented herein allow in-
   cooperating IP payloads, exercising multi-paths, verifying multicast
   trees, locating end stations, virtual segments and diagnosing
   connectivity problems. ICMP protocol is proposed as framework for
   error reporting. Document also presents network wide health
   monitoring, distribution and reporting methods that are intended for
   efficient troubleshooting.

Table of Contents

1. Introduction

   TRILL protocol has revolutionized how Layer 2 networks are being
   built and used. Legacy Ethernet networks provide single path for
   forwarding traffic and require all of the switches in the network to
   learn end-station MAC addresses. TRILL, on the other hand utilize
   multiple active links for forwarding thereby maximizing the overall
   network bandwidth utilization. TRILL is simple plug-and-play
   solution and does not require intermediate devices to learn MAC
   addresses of end-stations. These powerful characteristics of TRILL
   optimize performance and increase scaling limits. However, with that
   comes increased difficulty in diagnosing connectivity problems and
   locating end stations.

   Network operators are used to troubleshooting legacy networks with
   single paths. Legacy devices maintain forwarding database of all
   end-station addresses in the Layer 2 network. Network administrators
   can trace the path taken by specific MAC address by examining the
   forwarding databases of devices. TRILL core switches, by design do
   not maintain end-station address database. Hence, administrators may
   not be able to trace a path taken by a specific MAC address by
   tracing the forwarding databases. Additionally, a given device may
   utilize multiple active paths to reach to a destination and may use
   a completely different forwarding topology for multicast traffic
   than it would use for unicast traffic. These challenges mandate the
   presence of an effective tool set to monitor and diagnose data plane
   failures in TRILL networks. These tools and protocols must stay as
   close as possible to the forwarding paths taken by actual data. OAM
   frames should not leak to end stations or out of the TRILL network
   to legacy networks.

   TRILL base protocol specification [RFC6325] does not specify
   algorithm for selecting a path from a set of equal cost paths to
   forward a given flow. The majority of traffic in the networks is IP
   centric and most devices deploy some sort of hashing algorithm to
   identify the forwarding path from set of equal cost paths for a

given flow. Thus, it is desirable to use IP address and TCP/UDP port information as inputs to the ECMP selection hash function. Use of such higher level information provides better distribution of flows across multiple equal cost paths. This document, propose a framework that allow specifying, various combinations of payloads including IP payloads and actual payloads.

As TRILL based networks get deployed, during the transition period, it may be required for TRILL RBridges to co-exist with legacy networks. It is very helpful for the network operator if TRILL data plane failure detection tools allow isolating problem to specific legacy device or at least to the interface(s) that the downstream legacy device is connected. Solutions presented in this document facilitate identifying legacy devices or RBridge interfaces legacy devices are connected to.

ICMP (Internet Control Message Protocol)[RFC 792] has been in use for nearly three decades. ICMP multipart extensions [RFC4884], propose methods to extend ICMP messages to include additional information, without changing or inventing new ICMP message types. In this document we utilize ICMP for reporting of errors. ICMP multipart extensions will be utilized to define additional information that is specific to TRILL. Additionally use of ICMP allows sending error reports either in-band or out-of-band. Use of out-of-band ICMP allows network operators to diagnose uni-directional path failures easily. Also, the same ICMP infrastructure can be utilized to generate unsolicited error notifications for TRILL data plane failures, such as Destination unreachable, Time Exceed (TTL expiry), Parameter Mismatch (MTU mismatch) etc..

Availability of Network health information is a valuable starting point for any failure detection process. In this document we present the concept of network regions, monitoring of network regions and distribution of network health.

Diagnostic tools are also commonly referred to as OAM (Operations, Administration and Maintenance). In this document we use words diagnostics and OAM interchangeably. Unless explicitly specified both the words means the same.

1.1. Motivation

Currently published TRILL OAM solutions, [TRILLCH] and [TRILLOAM], mainly focus on data plane encoding and individual tools. The encoding methods presented in [TRILLCH] and [TRILLOAM], require defining OAM channel that utilize a special EtherType. Implementations that utilize ECMP selection algorithms based on

higher layer address information may require flexible OAM channel
that allow specifying different payloads including IP based
payloads.

Availability of network health information is important for
efficient isolation of network connectivity problems. Currently
there are neither standard sets of such data to be distributed nor
framework to distribute network health data. Lack of such leads to
cumbersome and time consuming troubleshooting of network
connectivity issues, especially in multi-vendor networks.

Device virtualization is an increasing trend in datacenters and
large enterprises. Physical servers may host multiple virtual
servers and these virtual servers may move from physical server to
physical server based on load balancing policies. As part of network
connectivity problem isolation, it is important to identify the
location of the virtual servers and RBridges they are connected to.
Currently, administrators are required to utilize multiple tools to
locate these virtual machines and connecting RBridges.

ICMP has been in use over three decades as the primary OAM tool of
IP infrastructure. It is highly desirable to utilize the framework
of existing infrastructure such as ICMP, thereby leveraging
knowledge, implementation and time to market.

TRILL networks can co-exist with multi access LAN networks at the
boundary of the TRILL network. TRILL protocol [RFC6325], introduced
Designated RBridge (DRB) and Appointed Forwarder (AF) concepts to
ensure loop free forwarding and load splitting at the boundary of
TRILL and multi access LAN networks. Discovery of DRB,AF and
associated VLANs are important for effective fault isolation at the
TRILL and multi access LAN boundary. Currently there are no known
tools available for the purpose.

In this document we propose a framework and solution suite that will
address the above.

1.2. Contributors

Many people contributed with ideas and comments. Among all,
following people made notable contributions to all parts of this
document and spend time reviewing, debating and commenting to ensure
this specification addressees the problem space.

Ian Cox, Ronak Desai, Satya Dillikar ,Rohit Watve, Ashok Ganesan and
Leonard Tracy.

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].

In this document, these words will appear with that interpretation
only when in ALL CAPS. Lower case uses of these words are not to be
interpreted as carrying RFC-2119 significance.

3. Protocol Architecture Overview

Effective OAM solution is not only a set of tools but a wholesome
solution that covers all aspects of OAM, such as tools, monitoring,
reporting etc. Solution presented in this document contains multiple
subcomponents that cover various elements of the total solution.
There are six subcomponents in the proposed architecture. These
subcomponents collectively are called TRILL OAM Protocol. Here we
present an overview of the architecture of the solution and explain
the purpose of each of subcomponents and interaction between
different subcomponents. Subsequent sections cover details of each
of the subcomponents.

```
+--+
|  | +------------+      +----------+
|  | | Reporting  |      |          |
|  | +------------+      |          |
|  |       ^             |  ISIS    |
|T |       |             |          |
|R | +------------+      | (GenApp) |
|I | |            |<-->  |          |
|L | | Distribution|     |          |
|L | +------------+      +----------+
|  |       ^
|O |       |
|A | +------------+
|M | |            |
|  | | Monitoring |
|P | +------------+
|r |       ^
|o |       |
|t |       v             +------------+
|o | +-----------+       | TRILL      |
|c | |           |       | Data Plane |
|o | | Tools     |<-->   |            |
|l | +-----------+       +------------+
|  |
+--+
```
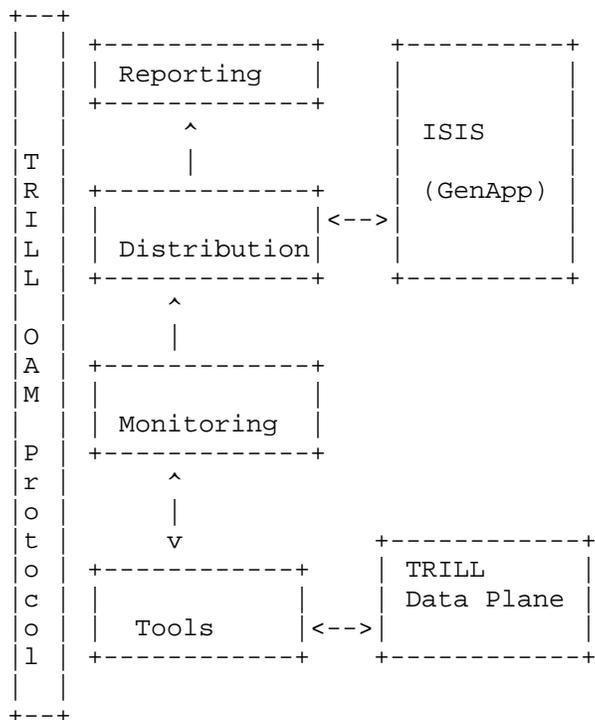
                    Figure 1 Architecture Overview

3.1. Overview of Tools

   The Tools subcomponent consists of series of utilities to implement
   various data plane monitoring and failure detections methods.
   Individual tools are invoked directly by the user or by the
   monitoring subcomponent. Individual tools allow, where applicable,
   for callers to specify options such as ECMP coverage, destination
   RBridge nickname, pay-load etc. Tools interface with the TRILL data
   plane layer to send and receive OAM frames. At the time of writing
   following tools are included as part of the tool set.

      1. Loopback Message (Ping)

      2. Path Trace Message (Trace route)

      3. Multicast Tree Verification (mtv)

   4. MAC discovery

   5. Address Binding Verification

   6. IP End-station Locator

   7. DRB-AF discovery

   8. Notification messages

Tools, based on their intended use, can be classified in to 3
broader categories as below.

```
+-------------------+----------------------------+
| Category          | Tools                      |
+-------------------+----------------------------+
| Fault Verification|   Loop Back Message        |
+-------------------+----------------------------+
| Fault Isolation   | Path Trace Message,        |
|                   | Multicast Tree Verification|
+-------------------+----------------------------+
| Auxiliary         | MAC discovery              |
|                   | Address Binding Verification|
|                   | IP End-station Locator     |
|                   | DRB-AF Discovery           |
|                   | Error Notification         |
+-------------------+----------------------------+
```
3.2. TRILL Data Plane

The TRILL data plane receives and transmits frames on behalf of the
tools subcomponent. As far as the encapsulation is concern, TRILL
data plane layer treat these frames exactly as it would treat a
regular data frame. In fact one of the key design goals is to
maintain TRILL data plane diagnostic (OAM) frames as close as
possible to actual data frames. Additionally, implementation MUST
satisfy the following requirements:

   1. OAM frames SHOULD NOT leak in to legacy Ethernet or to end
      stations outside the TRILL cloud

   2. RBridge MUST have ability to identify OAM (diagnostics) frames
      intended for a destination RBridge.

3. RBridgeS SHOULD have ability to identify TRILL data OAM frames
   that are not intended for itself and forward such frames
   without assistance from the CPU.

We explain in Section 6 various methods available to identify
TRILL OAM (diagnostic) frames intended for the local RBridge and
satisfy above requirements.

3.3. Monitoring

The Monitoring subcomponent utilize the tools subcomponent to
monitor the TRILL data plane and proactively detect connectivity
faults, configuration errors (cross connect errors) etc. The
monitoring subcomponent provides options to specify frequency,
retransmission count, ECMP choice and all other applicable options
to the specific tool being used to implement the monitoring service.
Based on the configuration specified by the user, the monitoring
subcomponent periodically invokes the applicable tools.
Additionally, based on configuration, monitoring results are
propagated to the distribution subcomponent. Monitoring results are
always associated with a monitoring region. The monitoring region is
an administrative partition of the network such that it: 1. Maximize
the fault coverage, 2. Optimize network health data summarization.
More details of regions are discussed in Section 10.

3.4. Distribution

The distribution subcomponent has two primary inputs

   o Data from the Monitoring Layer

   o Data from other RBridges via ISIS GenApp

The distribution subcomponent performs the following functions:

   o Advertising locally generated data

   o Applying Advertising policies and re-advertising received data

   o Maintaining the network health Database

Details of distribution layer and data handling are presented in
section 10.

3.5. ISIS

   TRILL OAM protocol suite proposed in this document utilize ISIS to
   distribute

   OAM capability of individual RBridge

   In-band OAM IP and MAC address

   Above, OAM capability and In-band OAM address information are
   advertised using ISIS MT-Protocol extensions.[section 7.]

   Network monitoring data are distributed using ISIS GenApp extension
   methods specified in [GenApp]. Details of encoding and proposed TLV
   definitions are defined in detail in section 7.

3.6. Reporting

   The Reporting subcomponent allows users to define and use various
   reports on network health. The Reporting subcomponent utilize data
   available in the distribution subcomponent to generate requested
   reports. Sample reports are listed in Appendix A.

4. Frame Format

   TRILL data plane diagnostic (OAM) frames can be broadly classified
   in to three types: request, response and notification. Request
   messages are generated to measure TRILL data plane characteristics,
   such as connectivity. Response messages are generated by a RBridge
   in response to a request. Notifications are unsolicited messages
   generated due to certain failures such as unreachable destination.
   Details of individual messages are covered in later sections. Here
   we present frame encoding format for Request, Response and
   Notification messages.

4.1. Encoding of Request message

```
        +-----------------------------+
        | TRILL outer Header          |
        +-----------------------------+
        | TRILL Header                |
        +-----------------------------+ ---
        | L2 Header + EthType         |  ^
        +-----------------------------+  | Diagnostic
        | IP Header (including TCP/UDP)|  | payload
        +-----------------------------+  |
        | User defined data or        |  |
        . padded to zero              .  |
        |                             |  v
        +-----------------------------+ ---
        | ICMP Header                 |
        +-----------------------------+
        |ICMP extensions (optional)   |
        +-----------------------------+
```
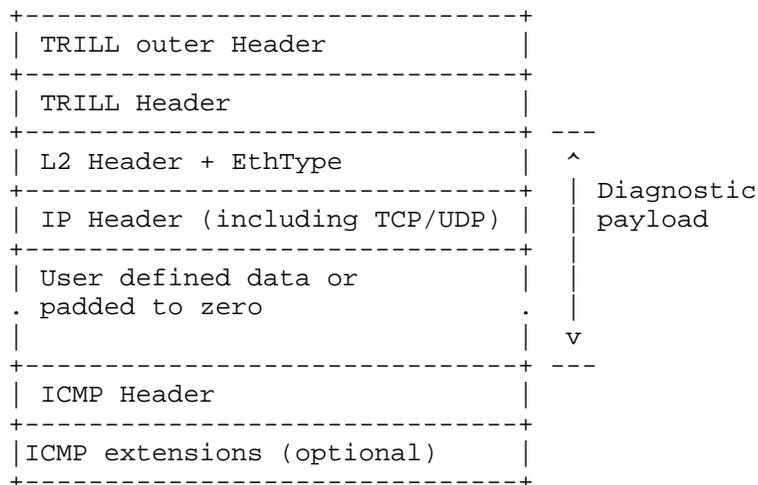
        Figure 2 Encoding TRILL data plane diagnostic request message

   The above diagram depicts encapsulation of TRILL data plane
   diagnostic request frames. Encoded in the frame is the diagnostic
   payload. The diagnostic payload is a flexible structure that allow
   user to specify different kinds of payloads, including actual
   payloads. Most hardware implementations use
   IPDA:IPSA:DestPor:SrcPort based hash method to select ECMP paths for
   IP frames. For non IP payloads, RBridges normally uses a Layer 2 MAC
   DA and SA based hash for selecting an ECMP path. Flexible diagnostic
   payload allow user to drive end to end ECMP selection based on
   payload without needing additional hardware. Also, in terms of
   forwarding, this keeps diagnostic frame as close as possible to data
   frames. The length of the diagnostic payload must be deterministic.
   We propose a fixed 128 byte size for the diagnostic payload section
   of the OAM frame. This allows including IPv6 frames with multiple
   802.1Q tags in to the diagnostic payload. The remaining bytes are
   set to zero, if the specified frame is smaller than the 128 byte
   fixed size.

   ICMP header immediately follows the diagnostic payload. The ICMP
   header is constructed as defined in [RFC792].

   ICMP multi part extensions [RFC 4884] are defined to carry
   additional information and are encoded after the ICMP
   header.[section 8.]

4.2. Encoding of Response Message

```
+-----------------------------+
| IP Header                   |
+-----------------------------+
| ICMP Header                 |
+-----------------------------+
|                             |
|   original frame            |
. (TRILL Header +             .
.  diagnostic payload)        .
|                             |
+-----------------------------+
|ICMP extensions              |
+-----------------------------+
```
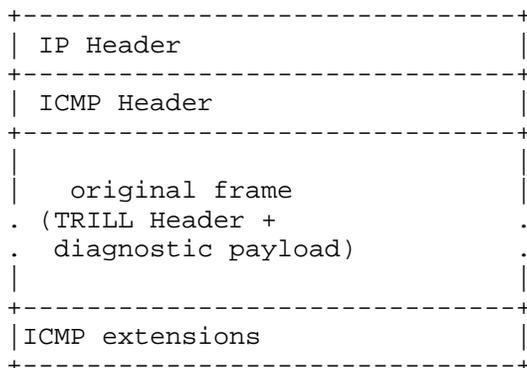

          Figure 3 Encoding of OAM response message

   The above diagram depicts encoding of OAM response messages. If in-
   band delivery is requested, the OAM response message MUST be encoded
   as payload in a TRILL data frame. The ingress RBridge nickname MUST
   be set to the RBridge nickname of the node generating the response.
   Egress RBRidge nickname MUST be set to the ingress RBridge nickname
   of the, original TRILL data frame that triggered this response.

    Normal IP forwarding rules MUST be followed, if an out-of-band
   response is requested.

4.3. Encoding of Notification Message

    Notification messages are generated in response to an error
   condition such as delivery failure due to incompatible MTU or
   destination RBridge not in the forwarding table etc.. Out-of-band
   responses are generally indicated by explicitly including the
   indication to receive an out-of-band response in the TRILL OAM
   request frame. Since notifications are generated in response to
   regular data frames, the originator RBridge may not have methods to
   identify IP address required to deliver an out-of-band response.
   Hence, in this document we propose to deliver Notification messages
   in-band. Delivery of out-of-band messages are outside the scope of
   this document.

   The RBridge generating the Notification message MUST include up to
   128bytes of the original frame that triggered the notification
   message. If the original frame contains less than 128 bytes, then
   the remaining bytes MUST be padded with zeros.

```
         +-----------------------------+
         | IP Header                   |
         +-----------------------------+
         | ICMP Header                 |
         +-----------------------------+
         |                             |
         |    original frame           |
         . (TRILL Header + L2+ Ethtype  .
         .  + data)                     .
         |                             |
         +-----------------------------+
         |ICMP extensions              |
         +-----------------------------+
```
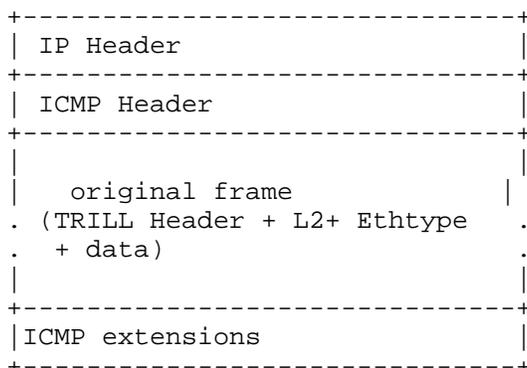
Figure 4 Encoding of Notification message

The TRILL outer header of the frame that triggered the notification
message is not included in the notification message. The Next-Hop
header information in the original frame is of local significance to
the specific link and may not be of interest to the originator of
the data frame.

The Following error messages are currently supported

   o Time Expiry
   o Destination Unreachable
   o Parameter Problem

Additional TRILL OAM error codes may be specified as ICMP multipart
extensions in above notifications messages. These error codes
indicate the cause of the error. Please see section 8. for error
code definitions and section 9.8. for theory of operation.

5. 127/8 in-band OAM IP address

   In this document we propose to use same ICMP framework deployed in
   IP infrastructure for communicating OAM information. RBridges are
   not required to have IP interfaces enabled. However, in order to
   receive and process ICMP messages, RBridges are required to have at
   least a pseudo IP address. In this document, we propose to use 127/8
   addressing scheme similar to the MPLS data plane failure detection
   methods [RFC 4373]. It is important that each RBridge have a
   straightforward method of identifying corresponding in-band OAM IP
   address of any given RBridge, without additional processing or
   lookups.

The 127/8 Address range is allocated for internal loopback addresses
[RFC 1122] and required not to be routed. RFC 4373 updates RFC 1122
to utilize 127/8 addressing to communicate between devices in a
peer-to-peer model that does not require routing. In this document,
we propose to use 127/8 addressing model to identify in-band IP
address required for OAM purposes. Additional methods are provided
as ISIS LSP extension to announce, other addresses, user may desire
to use for OAM in-band purpose. By default all RBridges MUST support
the 127/8 addressing model specified here.

Each RBridge nickname is 16bits wide [RFC6325]. Let's assume RBridge
nickname RB is divided in RB(msb) and RB(lsb), such that, RB(msb)
takes the upper 8bits of the RB and RB(lsb) takes the lower 8bits of
the RB. Corresponding in-band IP address of RB is
127.RB(msb).RB(lsb).100. Implementation MUST facilitate methods to
avoid conflicts between in-band OAM address and implementation
specific 127/8 address allocations.

5.1. IPv6 default in-band address

IPv6 based systems have two options to derive the in-band IP
address. The systems may choose, IPv6 native loopback address
::RBid:100 or IPv4 mapped IPv6 addressing format of
::FFFF:127.RB(msb).RB(lsb).100.

RFC 4379, MPLS Data Plane failure detection methods, utilize IPV4
mapped IPv6 addressing. One of the design objectives of the proposal
is to re-use as many existing OAM extensions as possible. Hence,
implementation that support IPv6 MUST utilize the IPv4 mapped IPv6
addressing fomat for default IPv6 in-band address. Deployments that
desire to utilize native addressing MAY advertise native IPv6 in-
band address using OAM extensions in section 7.

6. Identification of Diagnostic frames

In this document we have proposed to use the TRILL header as defined
in [RFC6325], without modifications. The standard TRILL header
currently, does not provide option to identify diagnostic frames.
Hence, it is important to have circumstantial methods to identify
diagnostic frames intended for the local RBridge and prevent leaking
of diagnostic frames outside of TRILL network. In this section we
explain, various methods to attain the above goals.

6.1. Identification of Layer 2 Flow

As stated earlier, most RBridges use Destination and Source MAC
address, combination to determine the next hop ECMP interface to

forward non IP frames. It is required to provide flexibility for the
user to specify destination MAC address and source MAC address. We
propose to use special EthType (TBD) to indentify OAM (diagnostic)
frames that contain non IP diagnostic payloads.

Each RBridge, if TRILL data plane OAM enabled, MUST provide
following processing:

   o Forward frames that have egress RBridge nickname equal to local
      RBridge nickname and EthType equal to Diagnostic Ethtype, to
      the Central Processing Unit (CPU). Such frames SHOULD NOT
      egress out of the RBridge.
   o The RBridge SHOULD not egress frames with Diagnostic Ethtype to
      non TRILL interfaces.

6.2. Identification of IP Flows

   As stated earlier, most RBridges use combination of IP address and
   Layer 4 information such as UDP/TCP Port, to determine the next hop
   ECMP interface to forward IP frames. Hence, it is important to
   provide flexibility for users to specify destination IP addressing
   and payload information.

   In this section we propose several approaches to identify OAM
   (diagnostic) frames with IP payloads that are addressed to the local
   RBridge for processing

   Method 1:

   Use of Well know Destination MAC address:

   We propose to use a well known diagnostic MAC address (TBD-DMAC-1),
   as the Destination MAC address of the inner Layer 2 header.

   Each RBridge, if TRILL data plane diagnostic is enabled, MUST
   provide the following processing:

   o Forward frames which have egress RBridge nickname equal to the
      local RBridge nickname and Destination MAC address of the inner
      Layer 2 header equal to the Well Known Diagnostic MAC address
      (TBD-DMAC-1) to the Central Processing Unit (CPU). If RBridge
      nickname is not equal to the local RBridge nickname, frame MUST
      be forwarded normally.
   o RBridge SHOULD NOT egress frames with the Diagnostic MAC
      address (TBD-DMAC-1) as destination address to non TRILL
      interfaces.

Method 2:

Use of Well know Source MAC address:

We propose to use a well known source MAC address (TBD-SMAC-1), as
the source MAC address of the inner Layer 2 header.

Each RBridge, if TRILL data plane diagnostic is enabled, MUST
provide following processing:

  o Forward frames that have egress RBridge nickname equal to the
    local RBridge nickname and source MAC address of the inner
    Layer 2 header equal to Well Known source MAC address (TBD-
    SMAC-1), to the Central Processing Unit (CPU). If the egress
    RBridge nickname is not equal to the local RBridge nickname
    then the frame MUST be forwarded normally.
  o Each RBridge SHOULD NOT egress frames with Well known MAC
    address as source address to non TRILL interfaces.
  o RBridge SHOULD NOT dynamically learn the well known Source MAC
    address (TBD-SMAC-1) specified above.

Method 3:

Use of RBridge specific OAM MAC address:

Each RBridge may advertise, MAC address for the purpose of receiving
OAM frames with IP payloads. Sending RBridges may us the advertised
MAC address as the destination MAC address of the inner Layer 2
header of originating diagnostic request frames.

Each RBridge, if TRILL OAM is enabled MUST provide following
processing:

  o Forward frames that has egress RBridge equal to the local
    RBridge nickname AND Destination MAC address of the inner Layer
    2 header equal to the advertised RBridge specific OAM MAC
    address, to the Central Processing Unit (CPU).
  o RBridge SHOULD NOT egress frames with RBridge specific OAM MAC
    address as destination address to non TRILL interfaces.

6.3. Identification of Multicast Flows

Multicast frames are forwarded using one of the available multicast
trees in the TRILL network [RFC6325]. Selection of a multicast tree
is done at the ingress RBridge. Multicast frames are directed to a
selected multicast tree at the ingress. Hence exact payload
definition is not required for the purpose of ECMP selection.
However, based on multicast pruning, certain multicast addresses may
not be required to be forwarded to all members of the tree.
Intermediate switches perform, (S,G) or (*,G), forwarding based on
IP addresses for IP frames and MAC address for non IP frames. Hence,
in order to verify the effect of multicast pruning users may require
methods to specify Layer 2 and/or IP addressing information, as
applicable. There are two types of multicast tree verification
messages:

   o Overall Tree Verification Messages
   o Pruned Tree Verification Messages

6.3.1. Identification of overall tree verification frames

We propose to utilize a well known multicast diagnostic MAC address
(TBD-GMAC-1) for this purpose. If TRILL data plane diagnostics are
enabled, this specific MAC address MUST be installed on every
RBridge for all tress and MUST NOT be subject to pruning.

Each RBridge performs (*,G) forwarding of the frames based on the
well known multicast diagnostic MAC address (TBD-GMAC-1) in the
inner Layer 2 destination address. Additionally, it send a copy of
the frame to the CPU for analysis and generates a response to the
requester. Please see section 8.3 for details of multicast tree
verification message processing.

A RBridge SHOULD NOT egress multicast frames with above diagnostic
MAC address in to non TRILL interfaces. Also, RBridge MUST discard
any native frame received on non TRILL interfaces with the above
diagnostic MAC address as the destination MAC address.

6.3.2. Identification of Layer 2 Multicast group verification frames

We propose to utilize the diagnostic EthType (TBD) that was defined
earlier for identification of Layer 2 group verification frames.
User SHOULD have the ability specify destination MAC address, source
MAC Address, VLAN and payload data up to 128 octets.

Each RBridge, performs standard multicast forwarding. Additionally,
if EthType of the frame is equal to the well known diagnostic

Ethtype (TBD), the RBridge sends a copy of the frame to the CPU for analysis and generating response to the requester. Please see section 9.3 for details of multicast tree verification message processing.

RBridge MUST NOT t egress multicast frames with above EthType in to non TRILL interfaces. Also, RBridge MUST discard any native frame received on non TRILL interfaces with the above EthType.

6.3.3. Identification of IP Multicast group verification frames

We propose to use the well known MAC address (TBD-SMAC-1) defined in section 6.2 as the source MAC address. Users have flexibility to define, IP Address, VLAN and other payload data upto 128 octets. The Destination MAC address is derived based on the IP Multicast destination address.

RBridges perform (S,G) or (*,G) forwarding using the IP address information. Additionally, each RBridge send a copy of the frame to the CPU, if the source MAC address matches the well known MAC address defined here in.

RBridge MUST NOT egress multicast frames with above source MAC address to non TRILL interfaces. Also, each RBridge MUST discard any native frame received on a non TRILL interfaces with the above source MAC address.

RBridge MUST NOT dynamically learn the well known source MAC address specified here.

6.4. Default OAM flow Parameters

Parameters specified herein SHOULD be utilized as default parameters. Parameters specified under the Fixed category MUST not be changed based on user specification and MUST be followed exactly as specified below.

Flow type	Default Values	Fixed fields
Layer 2	DA= Well Known MAC SA= RBridge Interface MAC VLAN=native VLAN	EthType=OAM(TBD)
IPv4 OR IPv6	IP Address = in-band address IP Dest. Port = 3503 IP Src. Port = TBD DA = OAM MAC of egress RBridge SA =ingress RBr interface MAC VLAN=native VLAN	EthType=0x8000 OR EthType=0x86DD
Multicast Tree Verification	SA= RBridge Interface MAC VLAN=native VLAN	DA=Well Known Multicast MAC EthType=OAM(TBD)
Layer 2 Multicast	DA= Well Known MAC SA= RBridge Interface MAC VLAN=native VLAN	EthType=OAM(TBD)
IP Multicast	IP Dest Address = Default OAM MCast address IP Src. Address = in-band-address IP Dest. Port = 3503 IP Src. Port = TBD DA = OAM MAC of egress RBridge SA =ingress RBr interface MAC VLAN=native VLAN	EthType=0x8000 OR EthType=0x86DD

Figure 5 Default Parameters of Diagnostic(OAM) Payloads

7. ISIS Extensions

   A new ISIS subTLV definition is required to announce the following
   OAM related information:

     o OAM capability
     o OAM in-band IP address
     o OAM in-band MAC address

   We propose to define a single sub TLV structure within ROUETER-
   CAPABILITY ISIS TLV (242), to announce the above OAM information.

```
 +--------+
 | Type   |
 +--------+
 | Length |
 +--------+--------+
 |ver| Res |v|i|m|o|
 +----------------+
 | Sender nickname |
 +----------------+
 |                |
 | OAM MAC address |
 |                |
 +----------------+
 |                |
 | OAM in-band    |
 . IP address     .
 |                |
 +----------------+
```
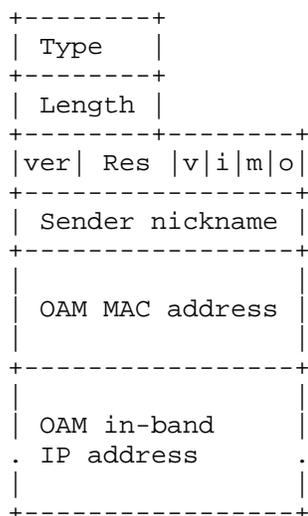
                  Figure 6 ISIS extension for OAM

   Type : (1 octet) TBD (one of the sub-TLV definitions under MT-PORT-
   CAP ISIS TLV)

   Length : (1 octet) Length of the subTLV, in octets, excluding Type
   and Length fields. Minimum 2.

   Ver : (4 bits) indicate the OAM version. Currently set to zero.

   Res : (1 octet), Reserved for future use. Set to zero on
   transmission and ignored on recipt.

   V   : (1 bit) if set, indicates IP address included in the TLV is
   IPv6. Only one of I or V bit MUST be set. If both are set, it is a
   malformed TLV and must be discarded without further processing.

I   : (1 bit) if set indicate IP address included in the TLV is
IPv4. Only one of I or V bits MUST be set. If both are set, it is
malformed TLV and must be discarded without any further processing.

M : (1 bit) If set, indicates MAC address is included in the TLV.

O : (1 bit) If set, indicates announcing RBridge is OAM capable.

MAC Address : (6 octets), IEEE MAC address, associated with the in-
band IP address. If included, the MAC address MUST precede the IP
address.

IP Address : (4 or 16 octets), OAM in-band IP address. If present
MUST follow MAC address.

Above PDU encoding MUST follow exact order as specified and fields
are not interchangeable.

NOTE: Both I and V flags MAY be set to zero to indicate that
announcing RBRidge desire to use the default OAM address. The
default OAM address is the 127/8 address derived as specified in
section 5.

8. ICMP multi part extensions

We propose to utilize a new Class-Num [RFC4884] to identify TRILL
OAM related extensions specified in this document and other related
documents. IANA has established a registry for ICMP extensions and
we intend to seek a Class-Num assigned for this purpose.

Within the TRILL OAM Class-Nume, C-Types are defined and registered
in the IANA to identify various different extensions specified
herein and other related future documents.

8.1. C-Type Definitions

Version: C-Type 1

Contain Version number, code and associated flags. Currently Out-of
band Request, Final and Cross Connect Error flags are defined.

```
 Bits
     31          24         16              2  1  0
     +-----------+---------+----------+--+--+--+
     | Version   | code    | Reserved |F |C | o|
     +-----------+---------+----------+--+--+--+
```
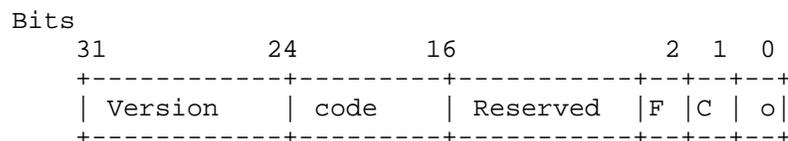
Figure 7 C-Type 1, Version and Flags

Version (8 bits): Currently set to zero

Code (1 octet) : TRILL OAM Message codes. See below for currently
available TRILL OAM Message codes.

Reserved (22 bits): Set to zero on transmission and ignored on
receipt

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

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

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

TRILL OAM Message codes:

0 : Loopback Message Request
1 : Loopback Message Response
2 : Path Trace Request
3 : Path Trace Response
4 : Time Expiry Notification (error)
5 : Parameter Problem Notification (error)
6 : Destination Unreachable (error)
7 : Multicast Tree Verification Request
8 : Multicast Tree Verification Response
9 : MAC Address discovery Request
10 : MAC Address discovery Response
11 : DRB discovery request
12 : DRB discovery response
13 : AF discovery request
14 : AF discovery response
15 : AF-VLAN discovery request
16 : AF-VLAN discovery response
17 - 255 : Reserved

Originator IP Address: (C-type 2)

Length of the ICMP extension header indicates whether the address is
IPv4 or IPv6. Please refer to RFC 4884 for ICMP extension encoding
and ICMP header structure.

```
Bits
      31                              0
      +------------------------------+
      |                              |
      . IP Address                   .
      |                              |
      +-----------+------------------+
```
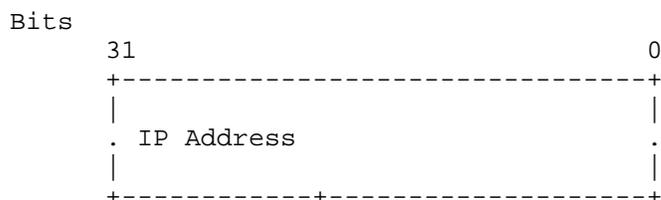
              Figure 8 C-Type 2 Originator IP address

Upstream Identification: (C-type 3)

The Upstream Identification C-type structure encodes upstream path
information such as upstream neighbor nickname, ingress interface
index (ifindex) and name of the ingress port.

```
 Bits
      31                              0
      +-----------------+--------------+
      |   nickname      | Reserved     |
      +-----------------+--------------.
      |          ifindex              |
      +-----------+-------------------+
      |   Slot          |    Port      |
      +-----------------+--------------|
      |    Speed        |    State     |
      +-----------------+--------------+
```
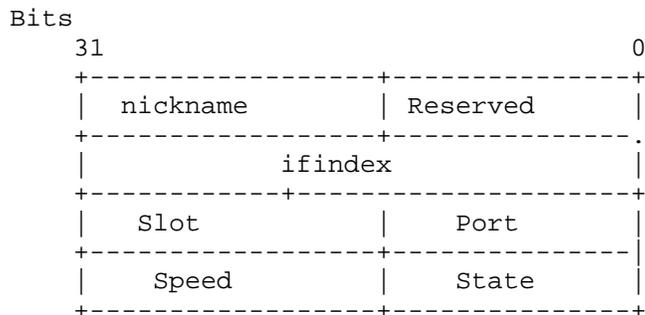
               Figure 9 C-Type 3 Upstream Identification

Nickname (2 octets): TRILL 16 bit nickname of the upstream RBRdige.
[RFCtrill]

Reserved (2 octetc)   : Reserved, set to zero on transmission
and ignored on receipt.

Ifindex (2 octets) : unsigned integer of local significance

Slot   (2 octets) : Slot number

Port   (2 octets) : Port number

Speed   (2 octets) : Speed in 100Mbps. Zero (0) indicates port
speeds less than 100Mbps.

State   (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Link Monitoring disable
All other values reserved.

Monitored VLAN(diagnostic VLAN) : (C-type 4)

Monitored VLAN c-type include in the ICMP extensions allows for
testing the integrity of the inner payload VLAN and the expected
VLAN. The expected VLAN is encoded in the Monitored VLAN c-type. The
destination RBRidge, compare the VLAN of the inner payload with the
VLAN value encoded in the Monitored VLAN c-type. If these two VLAN
values mismatch, RBRidge SHOULD set the cross connect flag in the
response. A RBridge MUST NOT set the cross connect error flag for
other than the above specified VLAN mismatch scenario.

```
 Bits
     16                              0
     +-------------------------------+
     | Reserved  | VLAN              |
     +-----------+-------------------+
```
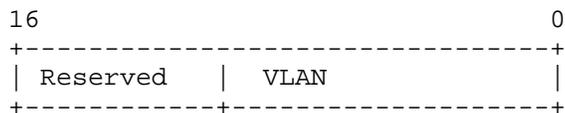
              Figure 10   C-Type 4 Diagnostic VLAN

Downstream Identification: (C-Type 5)

The Downstream Identification C-type carries multiple sets of data,
each corresponding to individual downstream neighbor among
collection of equal cost paths.

```
 Bits
   31                          0
     +-----------------+--------------+
     | ecmp count      | Reserved     |
     +-----------------+--------------+ ----
     | Reserved        | nickname     | ^
     +-----------------+--------------+ |
     |           ifindex              | | Next hop
     +-----------+--------------------+ | neighbor
     |  Slot           |   Port       | | information
     +-----------------+--------------| |
     |  Speed          |   State      | v ----
     +-----------------+--------------+ ----
     |                                |
     |  Repeat next hop neighbor      |
     .  identification for each       .
     |  neighbor                      |
     |                                |
     +--------------------------------+
```
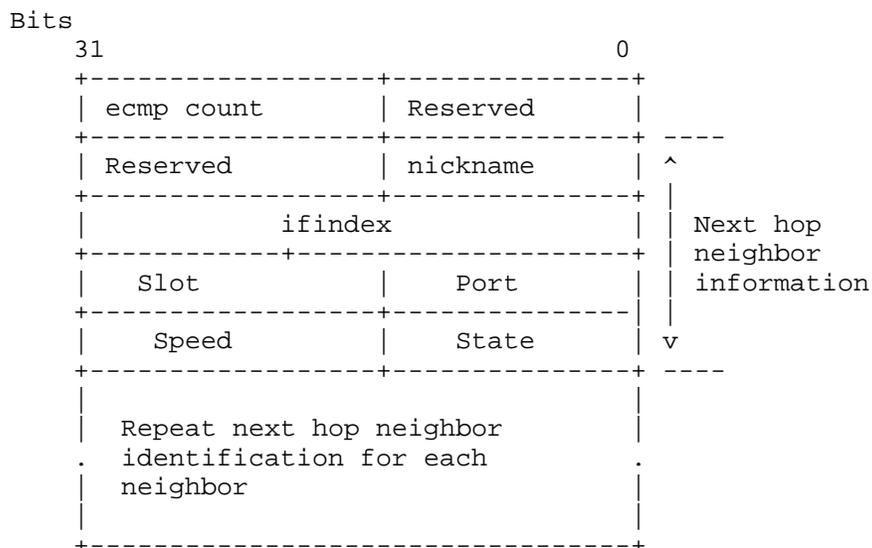
Figure 11   C-Type 5 Downstream Identification

Ecmp count (2 octets): Number of equal cost paths to the given
destination from this RBridge.

Reserved (4 octets): Reserved, set to zero on transmission and
ignored on receipt.

Next-hop neighbor information:

Nickname (16 bits): TRILL 16 bit nickname [RFCtrill]

Ifindex (2 octets) : unsigned integer of local significance

Slot   (2 octets) : Slot number

Port   (2 octets) : Port number

Speed   (2 octets) : Speed in 100Mbps. Zero (0) indicates port
speeds less than 100Mbps.

State   (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors

```
    4: Forwarding - errors
    5: Forwarding - oversubscribed
    6: Link monitoring disable
    All other values reserved.
```

    NOTE: Repeat Next-hop neighbor identification entry per each ECMP.
    Total number of neighbor entries MUST equal to ecmp count.
    Individual neighbor entry MAY have variable length.

    Path for this payload: (c-Type 6)

    Path for this payload indicates the next hop neighbor that this
    frame could have been forwarded on based on the payload hashing.

```
    Bits
        31                                 0
        +-----------------+---------------+
        |   nickname      |  Reserved     |
        +-----------------+---------------.
        |           ifindex               |
        +------------+--------------------+
        |   Slot          |    Port       |
        +-----------------+---------------|
        |   Speed         |    State      |
        +-----------------+---------------+
```
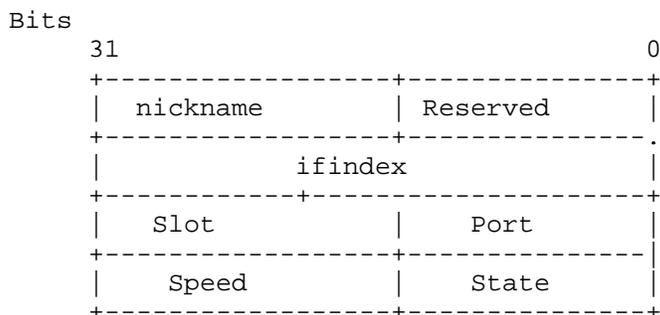
             Figure 12   C-Type 6 Path for this payload

   Nickname (16 bits): TRILL 16 bit nickname [RFCtrill]

   Ifindex (2 octets) : unsigned integer of local significance. 0xFFFF
   indicate CPU.

   Slot   (2 octets) : Slot number

   Port   (2 octets) : Port number

   Speed   (2 octets) : Speed in 100Mbps. Zero (0) indicates port
   speeds less than 100Mbps.

   State   (2 octets) : Represent the state of the port.

```
0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Link monitoring disable
All other values reserved.
```

DRB Information (c-Type 7)

```
    31              16        8      0
    +--------------+--------+--+-+
    | nickname     | state  | R|P|
    +--------------+--------+--+-+
```
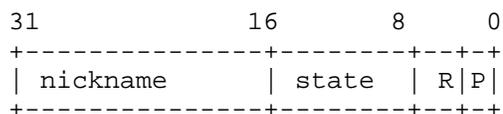
                    Figure 13   Nickname of the DRB

Nickname (2 octets) : TRILL nickname of the DRB

State   (1 octets) : DRB state

R   (7 bits) : set to zero on Transmission and ignored on
receipt

P   (1 bits) : Set when pseudo node bypass is indicated by
the DRB for the link

AF Information (C-Type 7)

Follow the same encoding as C-Type 6, above.

Nickname and state are of the AF.

Enable VLAN List (c-Type 8)

```
    31  27         16 12          0
    +--+----------+--+----------+
    |R | St-VLAN   |R | End-VLAN |
    +--+----------+--+----------+
```
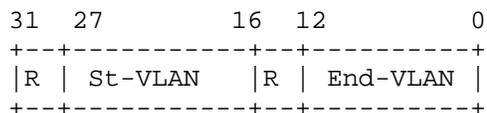
                    Figure 14   Enabled VLAN List

R (4 bits) : Reserved, set to zero on transmission and ignored on
receipt.

St-VLAN (12 bits) : Start VLAN

End-VLAN (12 bits) : End VLAN

Start VLAN and End VLAN represent the range of enabled VLANS. If the
VLAN range is non contiguous, then multiple Enabled VLAN lists MUST
be included, each representing a contiguous VLAN set.

Announcing VLAN set (c-Type 9)

Announcing VLAN list uses the same format as the Enable VLAN List
(c-Type 8)

```
   31  27         16  12           0
   +--+----------+--+----------+
   |R | St-VLAN  |R | End-VLAN |
   +--+----------+--+----------+
```
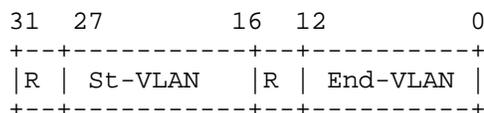

                 Figure 15   Announcing VLAN List

R (4 bits) : Reserved, set to zero on transmission and ignored on
receipt.

St-VLAN (12 bits) : Start VLAN

End-VLAN (12 bits) : End VLAN

Start VLAN and End VLAN represent the range of announcing VLANS. If
the VLAN range is non contiguous, then multiple of announcing VLAN
list MUST be included, each representing a contiguous VLAN set.

AF List (c-Type 10)

This c-Type lists the VLANs for which responding RBridge is a the
appointed forwarder.

```
 31  27         16 12          0
 +-------------+-------------+
 | Reserved    | nickname    |
 +--+----------+--+----------+
 |R | St-VLAN  |R | End-VLAN |
 +--+----------+--+----------+
```
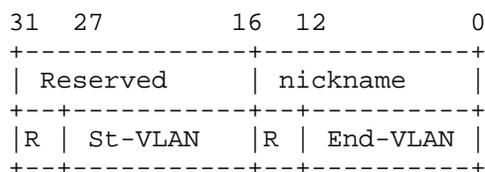
                    Figure 16   AF List

   Reserved (2 octets) : set to zero on transmission and ignored on
   receipt.

   Nickname (2 octets) : TRILL 16 bit nickname of the RBridge

   R (4 bits) : Reserved, set to zero on transmission and ignored on
   receipt.

   St-VLAN (12 bits) : Start VLAN

   End-VLAN (12 bits) : End VLAN

   AF List MUST be repeated for each of the contiguous VLAN ranges that
   the responding RBridge function as Appointed Forwarder.

   DRB Life Time (c-Type 11)

   DRB Life time indicates the Life time, of the DRB operational role,
   of the RBridge.

```
 31                                  0
 +-----------------------------------+
 |                                   |0
 +   Life Time                       +
 |                                   |1
 +-----------------------------------+
```

                    Figure 17   DRB Life Time

   Life Time (8 octets): Indicates the Life time of the operational
   role in seconds.

   AF Lifetime (C-Type 12)

   AF Life time indicates the Life time, of the AF operational role, of
   the RBridge for the specified VLAN.

Encoding follow the same format specified in C-Type 11.

Designated VLAN changes (C-Type 13)

Indicates number of times a given RBridge has observed Designated
VLAN changes. Each change may potentially lead to traffic
disruptions.

```
    15            0
    +-------------+
    | Change count|
    +-------------+
```

        Figure 18   Number of times Designated VLAN changes

Change count (2 octets): Indicates number of times a given RBridge
has observed Designated VLAN changes

RBridge scope List (c-Type 14)

```
  15           0
   +----------+
   |  R  | Nu |
   +----------+
   | nickname 1|
   +----------+
   .          .
   .          .
   | nickname n|
   +----------+
```

                Figure 19   Scope List c-Type 15

R (1 octet) : Reserved, zero on transmission and ignored on recipt.

Nu (1 octet) : number of nicknames listed

Nickname 1 .. n (2 octets) each: List TRILL RBridge nickname of in
scope RBridges.

Nicknames MUST be numerically sorted. With nickname1 the lowest to
nickname n the highest. This facilitate easy processing the
receiving RBridge.

Nu = 0 indicate no embedded nicknames in the message and response required from all RBridges, where applicable.

Multicast Tree downstream List (c-Type 15)

Multicast Tree downstream list provides information on downstream leaf Rbridges on the specified tree.

```
Bits
    31                              0
    +-----------------+--------------+
    | leaf count      | Reserved     |
    +-----------------+--------------+ ----
    | Reserved        | nickname     | ^
    +-----------------+--------------+ |
    |           ifindex              | | Downstream
    +------------+-------------------+ | leaf
    |   Slot     |       Port        | | information
    +-----------------+--------------| |
    |    Speed        |    State      | v
    +-----------------+--------------+ ----
    |                              |
    |  Repeat downstream           |
    .  leaf information for each    .
    |  downstream RBridge          |
    |                              |
    +-------------------------------+
```
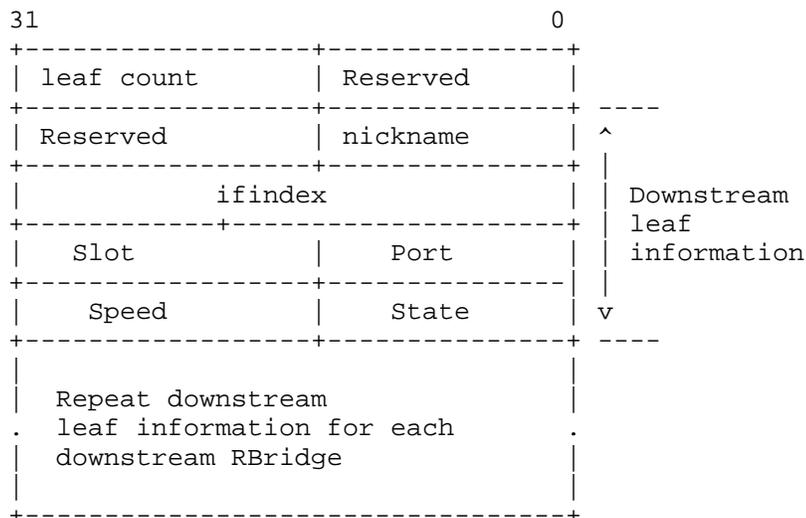
Figure 20   C-Type 5 Downstream Identification

Leaf count (16 bits): Number of RBridges downstream to this RBridge.

Downstream leaf information:

Nickname (16 bits): TRILL 16 bit nickname [RFCtrill]

Ifindex (32 bits) : Unsigned 32 bit integer that has only a local significance to the sending RBridge. Value 0xFFFF indicates CPU interface.

Slot   (2 octets) : Slot number

Port   (2 octets) : Port number

Speed   (2 octets) : Speed in 100Mbps. Zero (0) indicates port
speeds less than 100Mbps.

State   (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
All other values reserved.

NOTE: Repeat downstream RBridges reachability information per each
leaf node. Total number of neighbor entries MUST equal to leaf
count. Individual neighbor entry MAY have variable length.

MAC-discovery Address List (c-Type 16)

```
 15           0
 +-----------+
 | count     |
 +-----------+
 | MAC       |
 + Address 1 +
 |           |
 +           +
 |           |
 +-----------+
 |           |
 .           .
 .           .
 .MAC        .
 |Address n  |
 +-----------+
```

Figure 21   MAC-discovery Address List

Count (2 octets) : Number of MAC addresses embedded in the response

MAC Address (6 octets) : 6 octet MAC address

MAC-discovery response Address List (c-Type 17)

```
    15              0
    +-------------+
    | count       |
    +-------------+
    | T | VLAN    |
    +---+---------+
    |             |
    + MAC         +
    | Address     |
    +             +
    |             |
    +-------------+
    | Age         |
    +             +
    |             |
    +             +
    |             |
    +             +
    |             |
    +-------------+
    | Ifindex     |
    +             +
    |             |
    +-------------+
    | vNTAG       |
    +-------------+
    | Slot        |
    +-------------+
    | Port        |
    +-------------+
    | State       |
    +-------------+
    | Speed       |
    +-------------+
```
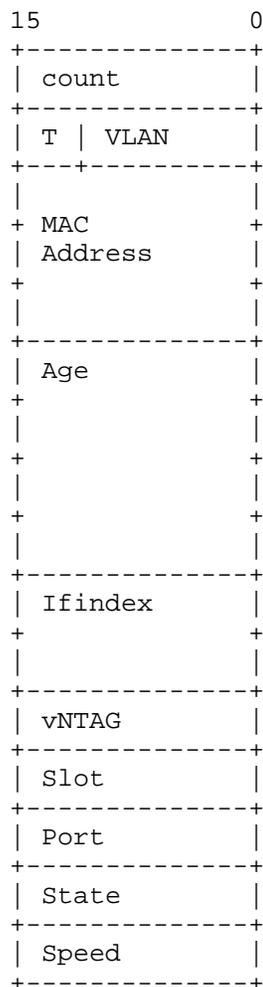
                    Figure 22   MAC-discovery response

   Count (2 octets) : Number of MAC addresses embedded in the response

   T   (4 bits) : Type of MAC address 0 - Dynamic, 1 Static, 2-15
   Reserved

   VLAN   (12 bits) : VLAN identifier associated with the MAC address

   MAC Address (6 octets) : 6 octet MAC address

Age (8 octets): Age of the MAC address in seconds. For a static MAC address, this field is ignored.

Ifindex (4 octets) : Interface index on which MAC address is learnt

Slot (2 octets) : Slot number of the interface on which this MAC address is learnt

Port (2 octets): Port number of the interface on which this MAC address is learnt.

vNTAG (2 octets): virtual TAG identifier associated with the MAC address. Value 0 indicate no vNTAG association with the MAC address.

Speed   (2 octets) : Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.

State   (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Un-monitored
All other values reserved.

Error code (c-Type 18)

Error code c-Type allows an RBridge to specify various error codes within high-level notification messages such as Time Expiry, Parameter Problem and Destination unreachable. The sub-error codes within each of the error code allow specifying further details of the error.

```
   Bits
      31                             0
      +-----------------+-------------+
      | Error Code      | sub-code    |
      +-----------------+-------------+
```
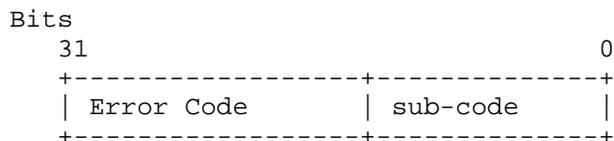
Figure 23   C-Type 18 Error code

Error Code (2 octets) : Identify the error. Currently following errors are defined

0 - VLAN non existent
1 - VLAN in suspended state
2 - Cross connect error
3 - Unknwon RBridge nickname
4 - Not AF
5 - MTU mismatch
6 - Interface not in forwarding state
7 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Sub-code (2 octets) : identify the sub-error code.
0 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Warning code (c-Type 19)

Warning code c-Type allow a RBridge to specify various error codes within high-level notification messages such as Time Expiry, Parameter Problem and Destination unreachable. The sub-warning codes within each of the warning codes allow to specify further details of the warning.

```
   Bits
     31                                 0
     +------------------+--------------+
     | Warning Code     | sub-code     |
     +------------------+--------------+
```
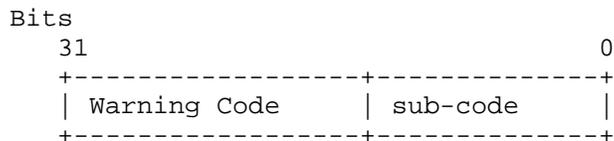
              Figure 24   C-Type 19 Warning code

Warning Code (2 octets) : Identify the Warning. Currently following Warnings are defined

0 - Inavlid RBridge nickname (RBridge nickname in the range 0xffco to 0xffff)
1 - Invalid VLAN (Reserved VLAN)
2 - AF VLAN list Mismatch
3 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Sub-code (2 octets) : identify the sub-error code.

0 - 0xFFFF - Reserved for future use and MUST not be used in
transmission.

Information code (c-Type 20)

Information code c-Type allow a RBridge to specify various
information codes within the high-level notification messages such
as Time Expiry, Parameter Problem and Destination unreachable. The
sub-info codes within each of the code allow specifying further
details of the information.

```
   Bits
       31                               0
       +------------------+--------------+
       | Information Code | sub-code     |
       +------------------+--------------+
```
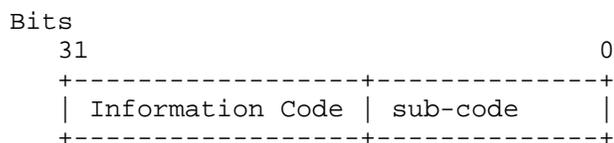
              Figure 25   C-Type 19 Information code

Information Code (2 octets) : Identify the Information. Currently
following Information are defined

0 - 0xFFFF - Reserved for future use and MUST not be used in
transmission.

Sub-code (2 octets) : identify the sub-error code.

0 - 0xFFFF - Reserved for future use and MUST not be used in
transmission.

9. Details of Diagnostic tools

In this section we present details of various diagnostic tools that
are identified as part of the solution. We assume, readers are
familiar with frame encoding methods, diagnostic frame
identification methods, and ISIS and ICMP extensions presented
earlier in the document. In this section we will only make reference
to the extensions and methods, please refer to prior section for
details.

9.1. Loopback Message

   Loopback message is utilized for fault verification. It verifies
   connectivity between two RBridges, for a specified flow. Monitoring
   subsystem may use Loopback Message for connectivity monitoring and
   proactive fault detection. Users may specify exact flow, part of it
   or not at all. Additionally, users may also specify, ECMP choice at
   the ingress. ECMP choice can be a specific index, set of index, all
   of the index or non. If no ECMP index specified, payload is used to
   determine the ECMP choice. Method of deriving the ECMP choice using
   payload is implementation dependent and is outside the scope of this
   document. However, CPU generating the Loopback message SHOULD use
   the same ECMP selection algorithm as the data plane. Additionally
   some implementation may allow users to specify the ingress interface
   that actual flow may ingress to the RBridge. Although ability to
   inject the data plane diagnostic frames from the ingress interface
   is optional feature, it is highly desirable, as it allows verifying
   end-end connectivity from an ingress port to an egress RBridge.

   Egress RBridge can sent its response either in-band or out-of-band.
   In-band-response, additionally allow to measure round trip delay.
   In-band responses are tagged with the same VLAN as the request
   frame. ICMP multi part extensions in the request message allow user
   to specify whether out-of-band response required. If out-of-band
   request required, IP address it desire to receive the response MUST
   be specified.

   Additionally, diagnostic VLAN, may be specified as part of the ICMP
   multi part extensions. Receiver RBridge may compare inner VLAN in
   the payload and the specified diagnostic VLAN. If the two specified
   VLAN values do not match, C flag in Version C-type SHOULD be set to
   indicate cross connect error..

9.1.1. Theory of Operation

9.1.1.1. Originator RBridge

   Identify the destination RBridge based on user specification or
   based on location of the specified address (see below sections for
   MAC discovery and address locator).

   Construct the diagnostic payload based on user specified parameters.
   Default parameters MUST be utilized for unspecified payload
   parameters. See Figure 5 for default parameters.

   Construct the ICMP Echo request header. Assign applicable
   identification number and sequence number for the request.

ICMP multi part extension Version MUST be included and set
appropriate flags. Specify the code as Loopback Message Request(0).

Construct following ICMP multipart extensions, where applicable

    o Out-of-band response request

    o Out-of-band IP address

    o Diagnostic VLAN

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

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

RBridge may continue to retransmit, the request at periodic
interval, until a response received or re-transmission count
expires. At each new re-transmission sequence number may be
incremented.

9.1.1.2. Intermediate RBridge

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

9.1.1.3. Destination RBridge

Destination RBridge performs, frame identification methods specified
in above section 5. If the Loopback message is addressed to the
local RBridge, then the RBridge forward the Loopback messages to the
CPU for processing. CPU performs frame validation and constructs the
response as stated below.

Construct the IP header for the ICMP echo response. If no out-of-
band response requested, IP address in the IP header MUST be in-band
IP address. If out-of-band response requested destination IP address
is the IP address specified in the request message. Source IP
address is derived based on the outgoing IP interface address.

Construct the ICMP echo reply header using the received ICMP echo
request.

Include the received TRILL header and diagnostic payload in to the
data field of the ICMP echo request frame [section 4.2.].

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

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

Error handling:

If VLAN cross connect error detected or inner.VLAN does not exsist
in the RBridge then generate Destination Unreachable message and
specify the cause using error codes.

9.2. Path Trace Message

Primary use of Path Trace Message, commonly known in the IP world as
"traceroute", is fault isolation. It may also be used for plotting
path taken from a given RBridge to another RBridge. Operation of
Path Trace message is identical to Loopback message except, that it
is first transmitted with a TRILL Hop count field value of 1.
Sending RBridge expect a Time Expiry message from the next hop or a
successful response. If a Time Expiry message is received as the
response, the originator RBridge record the information received
from intermediate node that generated the Time Expiry message and
resend the message by incrementing the previous Hope count value by
1. This process is continued until, a successful response is
received from the destination RBridge or Path Trace process timeout
occur.

9.2.1. Theory of Operation

9.2.1.1. Originator RBridge

Identify the destination RBridge based on user specification or
based on location of the specified address (see below sections for
MAC discovery and address locator).

Construct the diagnostic payload based on user specified parameters.
Default parameters MUST be utilized for unspecified payload
parameters. See Figure 4 for default parameters.

Construct the ICMP Echo request header. Assign applicable
identification number and sequence number for the request.

ICMP multi part extension Version MUST be included and set
appropriate flags. Set the code to Path Trace Request (2)

Construct following ICMP multipart extensions, where applicable

   o Out-of-band response request

   o Out-of-band IP address

   o Diagnostic VLAN

Specify the Hope Count of the TRILL data frame as 1 for the first
frame. Or use Hope Count value incremented by 1 if this is a
retransmission generated in response to received Time Expiry
message.

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

RBridge may continue to retransmit, the request at periodic
interval, until a response received or re-transmission count
expires. At each new re-transmission sequence number may be
incremented.

9.2.1.2. Intermediate RBridge

Intermediate RBridge receive the diagnostic frame as Hope count
expired frame. Based on flow encoding methods explained in above
section 5, RBridge identify TRILL data plane diagnostic frames from
actual data frames with Hope count expiry. Hop count time expiry
messages may be generated for actual data frames as well. However,
Hop count expiry message for actual data frames are always sent in-
band, as actual payload does not have methods to specify the
response delivery method.

CPU of intermediate RBridge that receives OAM frame with Hope count
expiry performs following:

Identify wheather in-band or out of band response requested.
Construct the IP header accordingly.

Construct the ICMP Time Expiry message as specified in RFC 792 and
RFC 4884. RFC 4884 specifies format of ICMP header when including
ICMP multipart messages.

Include original TRILL header and diagnostic payload of the original
frame as data for ICMP Time Expiry message. Update the length field
to reflect the size of the TRILL header and diagnostic payload.

Include following ICMP multipart extensions

Version

Set the code to Path Trace Response (3)

Nickname of the RBridge

Information of the ingress interface (speed,state,slot,port)

Index of the interface where frame was received

nickname of the upstream RBridge the frame was received

Downstream ecmp count

List of Downstream RBridges {nickname, interface index and interface information}

Downstream path this specific payload take { RBridge nickname, interface index and interface information}

Optionally include following ICMP multipart extensions

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

If in-band response was requested or the message was generated due to actual data frame, dispatch the frame to the TRILL data plane with request-originator nickname as the egress RBridge nickname.

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

9.2.1.3. Destination RBridge

Processing is identical to section 8.1.1.3

9.3. Multicast Tree Verification (MTV) Message

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

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

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

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

9.3.1. Theory of Operation

9.3.1.1. Originator RBridge

User is required at minimum to specify either the multicast trees that needed to be verified or Multicast MAC address and VLAN or VLAN and Multicast destination IP address. Alternatively, for more specific multicast flow verification, user MAY specify more information e.g. source MAC address, VLAN, Destination and Source IP addresses. Implementation, at minimum, must allow user to specify, choice of multicast trees, Destination Multicast MAC address and VLAN that needed to be verified. Although, it is not mandatory, it is highly desired to provide option to specify the scope.

Default parameters MUST be used for unspecified parameters. Please refer to Figure 5 for default payload parameters for MTV message.

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

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

Construct the diagnostic payload based on user specified parameters. For overall multicast tree verification message only multicast tree is specified as input. For generic multicast group verification, additional information such as group address is specified. Based on user provided parameters, implementation SHOULD identify whether the request is for overall multicast tree verification or for specific group verification.

For overall multicast tree verification, use well known multicast destination MAC address (TBD_GMAC-1) defined in above section 6.3.1. as the inner destination MAC address of the TRILL frame. Remaining parameters are derived based on default values specified in Figure 5

Construct ICMP echo request message header and include sequence number and identifier. Identifier and sequence number facilitate the originator to map the response to the correct request.

Version ICMP multipart extension MUST be included.

Code MUST be specified as Multicast Tree Verification Request (7)

Optionally, include following ICMP multipart extensions, where applicable

   o Out-of-band response request

   o Out-of-band IP address

   o Diagnostic VLAN

   o In scope RBridge list.

        o NOTE: Nu field in ICMP extension RBridge scope (section
          8.1.) MUST be set to zero, if response required from all
          RBridges.

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

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

RBridge may continue to retransmit, the request at a periodic interval, until a response received or re-transmission count expires. At each new re-transmission sequence number may be

incremented. At each re-transmission, RBRidge may further reduce the
scope to the RBRidges it has not received a response.

9.3.1.2. Intermediate RBridge

Intermediate RBridges identify multicast verification frames per the
procedure explained in section 6.3. .

CPU of the RBridge validate the frame and analyze the scope RBridge
list. If the local RBridge nickname is not specified in the scope
list, it will silently discard the frame. If the local RBridge is
specified in the scope list, RBridge proceed to 9.3.1.3 for further
processing.

9.3.1.3. In scope RBridges

RBridge go through following processing, upon identifying that it's
nickname is specified in the scope RBridge list.

Identify wheather in-band or out of band response requested.
Construct the IP header accordingly.

Construct the ICMP echo response message as specified in RFC 792.

Include TRILL header and diagnostic payload of the received OAM
message as data of the ICMP response message.

Include following ICMP multipart extensions

Version, update the code as Multicast Tree Verification Response (8)

Nickname of the RBridge

Name of the ingress interface frame was received

Interface index where frame was received

Nickname of the upstream RBRidge the frame was received

Downstream leaf node count

Leaf RBridge list {RBridge nickname, interface index and interface
name}

Optionally, if VLAN cross connect error detected, then set C flag
(cross connect error) in the versions extension.

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

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

Error Handling:

RBridge MUST generate applicable notification messages if any error
such as inner VLAN not available, detected against the OAM message.

9.4. MAC address discovery Message

MAC address discovery message is defined to discover following
information

   o RBridge nickname where the MAC address is learnt

   o Interface Index and Name on which the MAC address is learnt

   o Type (i.e. Static, Dynamic, Secure etc.)

   o Age of the MAC address

   o Virtual Interface Tag (vNTAG)

   o Interface Type (Legacy or TRILL Shared)

   o DRB on the VLAN (If Applicable)

   o AF for the VLAN (If Applicable)

   o Time AF operational (If Applicable)

  Optionally, an implementation may include the following information

   o System MAC address of the device connected to the port with
     which the MAC address is associated.

   o System information, such as name, IP address and location of
     the device connected to the port with which the MAC address is
     associated.

o Information related to this MAC address from the remote
  device..

The method of obtaining the above optional information is outside
the scope of this document. However, implementation may consider
link level control protocols such as LLDP for the purpose.

9.4.1. Theory of Operation

There are two possible options to implement MAC address discovery.
Either we may define a new MAC-discovery ISIS sub-TLV and use ESADI
to propagate the request (similar to the MAC-Reachability TLV
[RFC6165]) OR we may use multicast tree verification message and
include a ICMP multipart extension to indicate that the message is a
MAC discovery message.

Using the ISIS based method has disadvantage of being non real time
and subjected to protocol delays. The second method above is
independent of any control plane protocol implementation and can be
exercised in real-time. Hence, in this document, we propose to
utilize second method.

9.4.1.1. Originator RBridge

Use the well known Multicast MAC address described in section 6.3.1.
, above as the inner destination MAC address of the diagnostic
payload. Use the applicable source MAC address and VLAN. Use the
diagnostic EthType defined earlier as the EhtType. Pad the
remainder of the diagnostic payload with zero.

Construct ICMP echo request message and include sequence number and
identifier. The sequence number and identifier facilitate the
originator to map the response to the correct request.

Construct following ICMP multipart extensions

  o Version

  o Set the OAM code to the MAC address discovery request (9)

  o Indicate that this is a MAC discovery message

  o One or more MAC address to be discovered

  o VLAN ID of MAC addresses (optional)

  o Out-of-band response request (optional)

    o Out-of-band IP address (optional)

    o In scope RBridge list. If response required from all RBridges,
      then the Nu count in RBridge scope list MUST be set to zero.

Specify the TTL value of the TRILL data frame to the applicable
value.

Set the egress RBridge nickname to the nickname of the multicast
tree used for broadcast and unknown unciast.

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

An RBridge may continue to retransmit the request at periodic
interval until re-transmission count expires. At each new re-
transmission sequence number may be incremented. The RBridge scope
list of re-transmission messages MUST be pruned to include only the
response pending RBridges. It is possible that more than one RBridge
has learnt the requested MAC address. Hence the implementation MUST
wait until the total wait time expires and SHOULD NOT abort the
discovery process on receiving a single response.

9.4.1.2. Receiving RBridges

CPU of Intermediate RBridges receives a copy of the MAC discovery
frame through methods explained in section 6.3.2. and 6.3.1.

Receiving in scope RBridges analyze the embedded ICMP multipart
extensions to identify whether the request is for MAC discovery.

If the request is for MAC discovery, then the receiving RBridge
queries its forwarding database to identify, whether requested MAC
address is present with specified VLAN information.

The receiving RBridge generate responses only for identified MAC
entries. If there are no matching MAC entries, the receiving RBridge
silently discards the MAC discovery request.

If a matching MAC address is found, the receiving RBridge generates
a Destination unreachable ICMP message (Type = 3) and code = 12,
"Destination host unreachable for type of service". This essentially
indicates, it has found the MAC address but has reached the end of
the TRILL network where the MAC address is located.

RFC 4884 allow extension of ICMP messages. Only ICMP messages
Destination Unreachable, Time Expired and Parameter Problem are
currently extensible in RFC 4884 compliant manner. Other messages
are only extensible for known payload size and considered non
compliant to RFC 4884. For MAC discovery messages there is no
requirement to include original data payload. Also response to MAC
discovery can contain large amount of MAC address information.
Hence, we conclude to utilize Destination unreachable message as
opposed to using an ICMP echo response with fixed pay load size.

The receiving RBridge constructs the response as follows:

Construct the IP header based on the requested response type, in-
band or out-of-band. For an in-band response, use RBridge in-band IP
address. For an out-of-band response, use the provided egress
RBridge out-of-band address.

Construct the ICMP Destination Unreachable message per section 4.1
of RFC 4884. Specify, ICMP type=3 and code = 12. Specify the length
as zero. (i.e, no data included and ICMP extensions directly
follow).

Include the following ICMP multi part extensions;

    nickname of this RBridge. (This is required in the event of out -
of band response to identify the originating RBridge nickname)

    Version

    Code, set to MAC address discovery response (10)

Additionally, include the following ICMP multipart extensions, for
each MAC address that was specified in the request and is present in
the RBridge forwarding DB:

  o Interface Index and Interface Information
     (Speed,Slot,Port,State) on which MAC address learnt

  o Type (i.e. static, Dynamic, Secure etc.)

  o Age of the MAC address

  o Virtual Interface Identification (vNTAG)

  o Interface Type (Legacy or Trill Shared)

  o DRB on the VLAN (If Applicable)

    o AF for the VLAN (If Applicable)

    o Time AF operational (If Applicable)

   Optionally an implementation may include the following information:

     o The system MAC address of the device connected to the port with
       which the MAC address is associated.

     o System information, such as name, IP address and location of
       the device connected to the port on which MAC address is
       associated.

     o Information related to this MAC address from the remote device.

    If the response size is greater than the maximum MTU size of the
   outgoing interface, then multiple responses MAY be generated. The
   final response frame MUST contain ICMP multipart extension Version
   (C-Type 1) with F (final response)flag set.

   The response frame is delivered to the TRILL data plane for in-band-
   response.

   If out of band response was requested, the response frame is
   delivered to the IP protocol stack.

9.5. Address-Binding Verification Message

   Virtual machine provisioning is a very common practice in data
   centers and enterprises. It is normal for virtual machines to move
   from one physical machine to another physical machine. As a result
   ARP tables on gateways can be stale and network operators may need
   to resort to multiple tools to identify the location of a given IP
   address that is being diagnosed for connectivity. Even if the
   location of the server that host the given IP address is identified
   using other tools, additional steps may be required to further
   identify the RBridge that interface with the physical server.

   It is important to have set of tools that allow an operator to
   quickly and easily identify the physical MAC address associated with
   a given IP address, or IP addresses associated with a given physical

MAC address. Additionally, it may be required to identify the
RBridge that connects to the given IP address. In this section, we
present methods to identify MAC address to IP addresses or IP
address to MAC address bindings.

Address binding tools presented here need to be exercised from
either a router or an RBridge that has IP services enabled on a
given VLAN.

There are two different address binding resolutions required

   1. MAC address to IP addresses binding

   2. IP address to MAC address binding.

We propose to use invARP [RFC 2390] to resolve MAC address to IP
address(es) binding and ARP [RFC 826] to resolve IP address to MAC
binding information. It is possible a given physical server to host
multiple virtual machines (i.e. IP Addresses). Hence, it is expected
to receive one or more responses, to an invARP request. However,
invARP in its current form is incapable of identifying whether a
single multi-homed host or multiple virtual hosts. At the time of
RFC 2390 and original ARP standard RFC 892 were written, virtual
machine concept did not exist. Hence, these protocols in its current
form do not include virtual machine identifiers such as vNTAGs. This
lapse of identification of virtual machines, make troubleshooting of
large virtual machine networks, with dynamic server allocation, very
difficult. Hence, we propose to extend, ARP [RFC 892] and invARP
[RFC 2390], protocol to carry, virtual machine identification tags.

Upon discovery of MAC address or identification that a given MAC
address is associated with a valid IP addresses, user may employ the
locator utilities listed in section 9.6. to identify the
corresponding RBridge and associated interface information.
Alternatively, implementation may support ARP response snooping with
extension explained in 9.5.1 to encode RBridge and location
information into ARP or invARP responses.

9.5.1. Extension to ARP and invARP

RFC 2390 presents methods to discover protocol address associated
with a given hardware address. In this section we propose methods to
extend RFC 2390 and RFC 892 to encode additional virtual interface
tag information and device information that may facilitate
identifying physical machine locations.

It is important the extensions proposed in the standard are
transparent to current implementations.

Figure 27, below, depicts the format of an ARP/invARP frame with the
proposed extensions embedded.

ARP frame as defined in RFC 892 and RFC 2390 has a fixed structure
and include only the length fields for addresses. Implementations
index in to these fix address fields and do not check the total
length of the response frame as part of validation. Hence, we
propose to include the extensions at the end after the target
protocol address. Implementations that do not support the new
extensions will safely ignore these values.

We expect additional identification information carried in ARP and
invARP to be limited. Furthermore, these, identification information
have compact and deterministic size. Hence, we propose not to use
explicit, length identification field, instead derive the length of
the value field implicitly, based on the class and class types
defined below. ARP and invARP follow identical encoding structures.

```
 31                          0
+----------+-----------+
| Hw Addr  | Protocol  |
+-----+-----+-----------+
| HL  | AL  | Op Code   |
+-----+-----+-----------+
|                       |
.  Source Hw Address    .
|                       |
+----------------------+
|                       |
.  Source Proto Address .
|                       |
+----------------------+
|                       |
.  Target Hw Address    .
|                       |
+----------------------+
|                       |
.  Target Proto Address .
|                       |
+----------------------+
| Extensions            |
.                       .
|                       |
+----------------------+
```
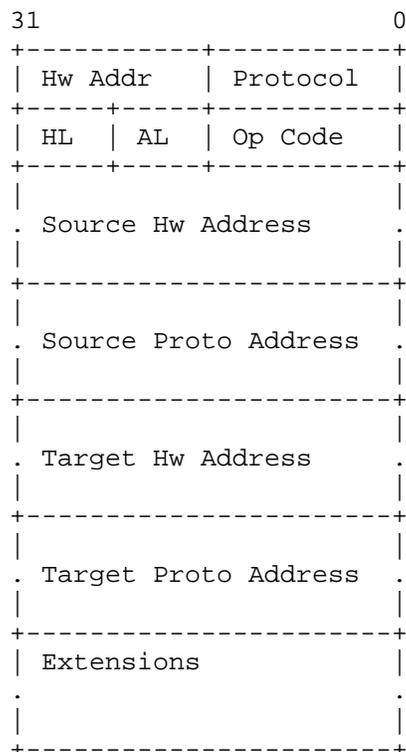
Figure 26   Encoding of ARP and invARP

9.5.1.1. Encoding ARP-invARP extensions

   ARP Extension encoding structure and proposed extensions are
   presented in this section. We propose a compact structure for ARP
   encoding. In Figure 27 "Class" identifies the Object Class and the
   "Class Type" (c-Type) within the class identify specific data
   element within the object class. C-Type implicitly indicates the
   size of the object. The encoded object size MUST NOT exceed the
   implied size of the corresponding Class and c-Type.

```
+-------+------+-------------+
|Class  |C-Type|             |
+-------+------+             +
|                            |
.       Object               .
|                            |
+----------------------------+
```
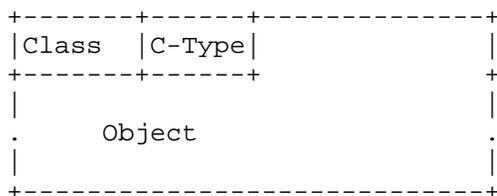
Figure 27   Encoding of ARP Extensions

Class : (1 octet). Define to identify the Object Class.

C-Type : (1 octet). Define Object type within Object class.

Object : (Variable octet, depends on the Class and C-Types)

```
+--------+--------+-------+-----------------------------+
|Class   |C-Type  |Name   | Description                 |
+--------+--------+-------+-----------------------------+
|  1     | 1      |vNTAG  |vNTAG of the interface       |
+--------+--------+-------+-----------------------------+
|  2     | 1      |RBridge|TRILL RBridge nickname       |
+--------+--------+-------+-----------------------------+
|        | 2      |ifindex|ifindex of RBridge interface ARP |
|        |        |       |response arrived             |
+--------+--------+-------+-----------------------------|
|        | 3      |Slot   |Slot id of RBridge interface ARP |
|        |        |       |response arrived             |
+-------------------------------------------------------+
|        | 4      |Port   |Port id of RBridge interface ARP|
|        |        |       |response arrived             |
+--------+--------+-------+-----------------------------+
```

Figure 28   Table of Class, C-Type and usage

Figure 28, above, presents Class, c-Type and application definitions.
vNTAG, rBridge, Slot and Port are each 2 octets in length. The length
of ifindex is 4 octets. All of the above extensions are optional. vNTAG
is inserted by the end station that is responding to the ARP request.
All other fields are inserted by the TRILL RBridge that interface with
the end-station and implement ARP response snooping. ARP response
snooping is similar to Dynamic ARP inspections, implemented by many
major vendors. Dynamic ARP inspection validates the Source IP address
of ARP response against known IP addresses to prevent ARP cache
poisoning by rogue stations. ARP response snooping, on the other hand,
intercepts ARP response frames and inserts required fields as defined

in this standard. Implementations may extend the dynamic ARP inspection framework to implement ARP response snooping.

In the interim, most end stations and servers may not insert the proposed vNTAG information. Hence, optionally, ARP response snooping, process on TRILL RBridge, MAY insert vNTAG information on behalf of the end station or server.

9.6. End-Station Attachment Point Discovery

   In traditional deployments, end stations and servers were relatively static in their locations. As a result localizing a fault was relatively easier.

   The virtual machine concept is an increasing trend in Datacenter and large enterprises. Dynamic load balancing policies of Virtual infrastructure, based on various load balancing policies, move virtual machines between different physical servers. This dynamic motion of virtual machines causes difficulty in associating a given virtual server to a RBridge. As a result, localizing a fault is a difficult task and requires use of multiple applications. Some virtual machine deployments utilize a single MAC address to represent all the virtual servers in a single physical server. Hence, it is important, to identify both the physical attachment point and the virtual segment information, such as VLAN and Virtual Tags.

   ARP/invARP extensions presented above facilitate discovery of the attachment information, however, some implementation may face scaling issues due to the large number of ARP requests. An alternative method is presented below.

   The End-Station attachment Point Discovery methods presented here, allow discovering, RBridge, interface information, VLAN, virtual Tags, etc, associated with a given IP Address.

   The End-Station attachment Point Discovery is a two step process. However implementations may present a single user interface that combines both the steps.

   Step 1: Utilize ARP to discover the MAC address associated with the specified IP address. Identify the ingress RBridge nickname by analyzing the TRILL header andidentify the VLAN information based on the inner VLAN.

   Step 2: Utilize MAC discovery methods explained above to discover, interface and virtual Tag information associated with the MAC

address discovered in above Step 1. Implementation SHOULD narrow the
scope of the MAC discovery to include only the RBridge and VLAN
discovered in step 1.

9.7. DRB and AF Discovery

The TRILL Base Protocol standard [RFC 6325] specifies support for
multi-access legacy network and shared segments between TRILL and
legacy devices. Legacy networks ensure loop free forwarding via the
IEEE 802.1D (Spanning Tree) protocol. RFC 6325 and RFC 6327 specify
loop prevention methods in mixed environments where the TRILL
network borders with a legacy multi-access network. RFC 6325 also
provide methods for load splitting of native traffic in to the TRILL
network. These are accomplished by having a single Designated
RBridge (DRB) for a given LAN segment which designates an Appointed
Forwarder (AF) for each VLAN on the segment to ingress and egress
traffic originating and destined to and from the legacy network.

Based on network dynamics, configurations, and failures, DRB and/or
AF designation may change from time to time. Hence, discovery of DRB
and AF is very important to effectively troubleshoot network
connectivity problems that involve TRILL and legacy networks
connected via non P2P TRILL interfaces.

DRB-AF discovery message has three variations.

   1. All DRB discovery

   2. All AF discovery

   3. VLAN,AF discovery

Above messages are identified with a unique TRILL OAM message code
(section 8.).

DRB-AF discovery messages allow for identifying the following
parameters:

   o Nickname of the DRB

   o STP Root Bridge identifier

   o Up time of AF (if responder is the AF)

   o Up time of DRB (if Responder is DRB)

   o Enabled VLAN List

    o Announcing VLAN List

    o DRB State (If Responder is the DRB)

    o AF State (If Responder is AF)

    o Pseudo Node bypass (If the Responder is the DRB)

    o Number of times the Designated VLAN has changed

    o AF List (nickname,start VLAN,end VLAN)(If the Responder is DRB)

   The above parameters are encoded in to the response message via ICMP
   multipart extensions (section 8.)

9.7.1. Theory of Operation

   DRB-AF discovery message utilize same addressing and format as the
   MAC discovery message (Section 9.4.)

9.7.1.1. Originator RBridge

   Follow the steps specified in section 9.4.1.1. , with the following
   exceptions

   Specify the message as one of the DRB-AF messages.

   If the message is VLAN,AF discovery message, then include the
   interest VLAN list.

9.7.1.2. Receiving RBridge

   Follow the processing steps specified in section 9.4.1.2. with the
   following exceptions:

   If RBridge is in the scope list or All-RBridge scope is specified,
   then the RBridge processes the message as follows:

   If the message is DRB discovery message then the receiving RBridge
   include the following information:

     o Response code set to DRB discovery response (12)

     o Nickname of the DRB

     o Nickname of AF of the specified VLAN

o STP Root Bridge identifier

o DRB Life time

o Enabled VLAN List

o Announcing VLAN List

o DRB State

o Pseudo Node bypass

o Number of times Designated VLAN change

o AF List (nickname,start VLAN,end VLAN)

If the message is an AF discovery or VLAN, AF discovery message,
then the receiving RBridge first validate whether the RBbridge is
the AF for the specified VLAN list and include following
information:

o Response code set AF discover response (14) or AF-VLAN discover
  response (16)

o Nickname of the DRB

o Nickname of AF of the specified VLAN or AF VLAN-List if VLAN is
  not specified.

o STP Root Bridge identifier

o AF Life time (i.e. How long has been AF)

o Enabled VLAN List

o Announcing VLAN List

o AF State

o Number of times Designated VLAN change

If RBridge is not the AF for specified VLAN then include ERROR code
Not AF (4) (see Figure 23).

If RBridge is AF for only a subset of VLANs specified in the request then include WARINING "AF VLAN list Mismatch" (3) and include the VLAN list that the RBridge is functioning as AF. (Figure 24)

9.8. Notification Messages

Notification messages are generated either due to regular TRILL data frames or TRILL OAM frames. Implementation MUST not generate notification messages on notification messages.

There are 3 types of Notification messages:

   o Time Expiry
   o Destination Unreachable
   o Parameter Problem

Within these Notification messages, error, warning and information ICMP extensions may be included to identify the details of the notification message. Section 4.3. above covers details of encoding Notification messages, section 8.1. covers ICMP extensions.

Time expiry messages are generated when TRILL hope-count field reach to zero. If applicable, It may contain additional error, warning or information extensions.

Destination unreachable notification may be generated for following scenarios; additional scenarios may be added later.

   o Egress RBridge nickname unknown
   o Inner VLAN does not exist or suspended
   o Not the AF for inner VLAN

Parameter Problem notification may be generated for following scenarios; additional scenarios may be added later.

   o Invalid RBridge nickname (RBridge nickname is one of the
     reserved 0xFFC0 - 0xFFFF)
   o MTU mismatch
   o Invalid VLAN (Reserved VLANs)
   o Interface state is not forwarding

10. Monitoring and Reporting

   Proactive identification of data plane failures are important part of maintaining Service Level Agreements (SLA). In traditional Layer

2, networks, there is only a single active path to monitor and both
multicast and unicast traffic follow identical paths. With TRILL,
there are multiple active paths and unicast and multicast traffic
take potentially different paths, depending on the flow parameters.

TRILL deployment in a typical data center may have 10's of 1000 of
links and 100's of RBridges. In such an enviorement, there may be
large number of active paths between two end points. As an example,
assume a topology with 4 RBridges connected serially with 32 ECMP
links at each hop. In the stated example topology, there are
32x32x32=32768 possible paths. Monitoring all of the possible path
combinations is not scalable. However, skipping some combination of
paths leads to reduce coverage and hence reduced effectiveness of
monitoring data. Even if one was brave enough to monitor all of the
links, analyzing and diagnosing a problem is quite cumbersome due to
the large amount of data. In other words, there must be methods to
scale the problem and present information in a more concise manner
that is still effective.

In this document we propose to use the "region" concept to partition
the network in to logical sections. Regions are monitored
independently. Detailed sets of monitoring data are distributed
throughout the region. A summary set of monitoring data is
distributed throughout the network. Network operators can obtain a
network health snapshot of the entire network from any RBridge in
the network. Detailed health report of a given region can be
obtained from any RBridge in the region.

An RBridge associate itself with a region through its interfaces. A
given interface can belong to one and only one region. An RBridge
can have multiple interfaces belonging to different regions. Each
RBridge is responsible for collecting monitoring data, organizing
the data in to regions and advertising the data to its peers. Please
see section 10.2, Advertising Policy for details.

In theory a network topology can be any arbitrary graph. In
practices, however, it is some set of sub-graphs repeating to
construct the overall topology. Each sub-graphs or set of sub-graphs
can be considered a region for monitoring purpose. The manner in
which regions are partitioned is an administrative choice such that;

    1. Maximize the fault coverage.

    2. Optimize network health data summarization.

As an example consider a typical datacenter topology depicted in
Figure 10. Typical datacenter may have multiple Points of

Demarcation (POD)s connected with an aggregation layer. A POD can be
considered as a region and may be individually monitored.

```
            +o--o+          +o---o+         +o--o+
  +~~~~~~~~~~~|   |~~~~~~~~|    |~~~~~~|   |~~~~~~~~~~~~~~~+
  |          | RB |        | RB |      | RB |              |
  |          +o--o+        +o---o+     +o--o+              |
  |            |  |          |  |        |  |              |
  |                                                        |
  |                Region Rm                               |
  |                                                        |
  |     |  |   |  |                        |  |   |  |      |
  |    +o--o+ +o--o+                      +o--o+ +o--o+     |
  +~~~| RB |~~| RB |~~~~~~~~~~~~~~~~~~~~| RB |~~| RB |~~~~~~~~+
   +~|    |  |~~|    |~~+                +~|    |  |~~|    |~~+
   | +o--o+ +o--o+    |                 | +o--o+ +o--o+    |
   |  |  |    |  |     |                 |  |  |    |  |     |
   |                  |                 |                  |
   |                  |                 |                  |
   |   Region R1      |    . . .        |   Region Rn      |
   |                  |                 |                  |
   |                  |                 |                  |
   |  |  |    |  |     |                 |  |  |    |  |     |
   | +o--o+ +o--o+    |                 | +o--o+ +o--o+    |
   +~|    |  |~~|    |~~+                +~|    |  |~~|    |~~+
   | RB |   | RB |                       | RB |   |RB |
   +-o--+   +o.o.+                       +-o--+   +o.o.+
```
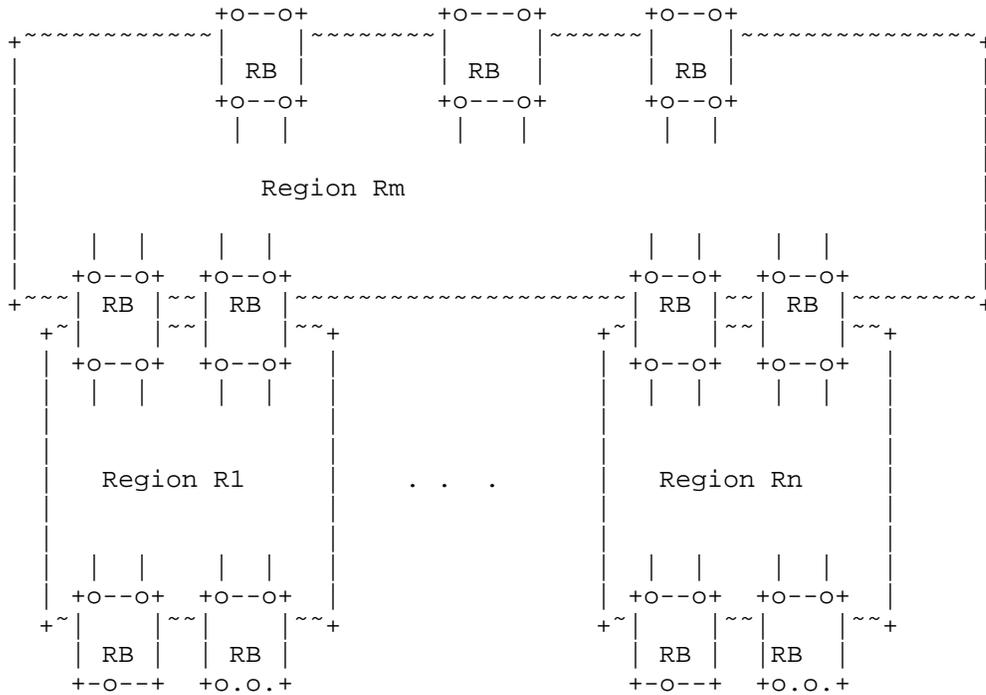
Figure 29   Example of "regions"

10.1. Data categories

There are 3 categories of monitoring data. They are, Summary
Category, Detail Category and Vendor Specific Category. The Summary
and Detail categories are mandatory. That is, every RBridge that is
compliant to this standard and support Monitoring, MUST support all
the elements defined under the Summary and Detail categories. The
Vendor specific Category is optional. Vendor specific data elements
are only available within the region. An RBridge that does not
understand the Vendor specific data elements forward them to
neighboring RBridges per Advertising Policies define in section
10.2. Individual data elements and structure of encoding Summary,
Detail and Vendor specific categories are presented in sections
10.3. – 10.5. .

10.2. Advertising Policy

    Each RBridge is responsible for advertising monitoring data to the
    OAM capable neighbors.

    Different interfaces on an RBridge can belong to different regions.
    However a given interface can belong to one and only one region. As
    a result a given RBridge may receive data from multiple regions.
    Each RBridge is responsible for advertising proper data categories
    over a given interface to the neighbor.

    Rule 1: No monitoring data are distributed:

        o On legacy interfaces

        o To neighbors not OAM capable

        o When ISIS state is not 2-way

        o When monitoring data advertisement is disabled

    Rule 2: Distribution of Summary category data:

        o Distribute on all OAM capable interfaces

        o Do not distribute summary data element of a region back to
          the originating region. (i.e. do not distribute on to
          interfaces that have the same region name as the data
          element)

        o Summary data for local region is derived from Detail data.
          (local summary data is never advertised into the local
          region per the above rule. However, it is advertised out to
          other regions the RBridge has interfaces in to)

    Rule 3: Distribution of Detail category

        o Distributed on OAM capable interfaces

        o Region of the data element and region of the interface must
          match for propagating a data element over an interface
          (i.e. Do not advertise to other regions)

        o Do not advertise data element back in to the originator
          RBridge.

Each RBridge distribute data at periodic intervals. Each RBridge
collects data it has received, analyzes them and redistribute
according to the rules specified above. The distribution interval
should be appropriately adjusted to not overload ISIS routing
operations.

Then Monitoring application is responsible for maintaining the
Application specific LSP. We propose to use Generic Application
Encoding methods explained in [GenAPP] for distributing Monitoring
data. TRILL operates in ISIS Level-1 layer, hence S,D flags defined
in [GenAPP] MUST be set to zero.

We propose to obtain specific Application ID [GenAPP][RFC5226] from
IANA for the purpose of registering TRILL Monitoring data
distribution.

Within the Application ID, context, a series of sub-TLV are defined
to carry specified information.

10.3. Summary Category

Then following individual data elements are defined within the
summary category.

   o Name of the region

   o Total number of RBridges in the regions

   o Total number of TRILL enabled ports in the region

   o Percentage of TRILL enabled ports down

   o Percentage of TRILL enabled ports oversubscribed

   o Maximum number of paths in the largest ECMP in the region

Then following structure encodes each of the data elements within
the summary category.

```
+----------+
| subTlv   |              2 octets
+----------+----------+
| Region-ID          |  4 octets
+--------------------+
|L |                 |
.--+                 .
.                    .
|                    |
+----------+---------+
| #Rbrdidge|            2 octets
+----------+
| #Ports   |            2 octets
+----------+
| #UpPorts |            2 octets
+----------+
| #OsubPort|            2 octets
+----------+
| #ErrPorts|            2 octets
+----------+
| #ECMP    |            2 octets
+----------+
| #DwnPorts|            2 octets
+----------+
```
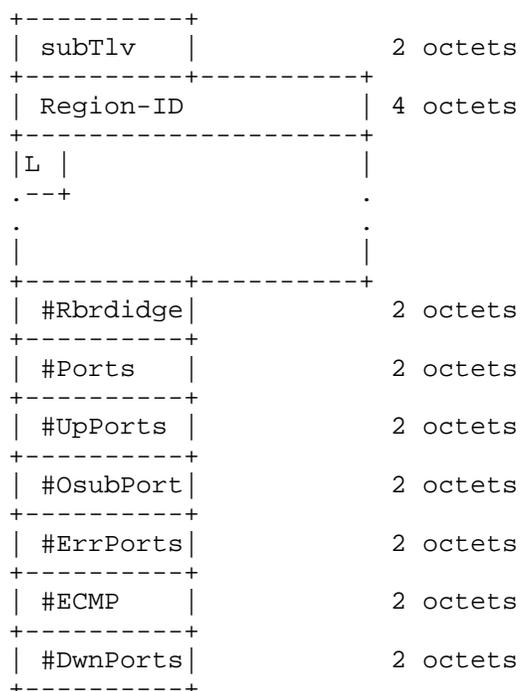
Figure 30   Encoding Summary Category Data

subType : (2 octets) is always 1 for summary category

Regiond-ID : (4 octets) is unsigned 32 bit integer identifier of the region

L   : (1 octet), length of the subsequent field

Region Name : '\0' terminated ASCII string of region name of variable size to maximum of 255 octets.

#Rbridge: (2 octets), number of RBridges in the region

#Ports: (2 octets) Total number of TRILL enabled ports available on this RBridge

#Up Ports: (2 octets) Total number of TRILL enabled ports that are operationally up.

#OSPorts : (2 octets) Total number of TRILL enabled ports that are oversubscribed.

#ErrPorts : (2 octets) Total number of TRILL enabled ports that are
indicating errors.

#DwnPorts : (2 octets) Total number of TRILL enabled ports that are
operationally down.

#ECMP : (2 octets) Maximum number of ECMP as seen by this region
ISIS routing table.

10.4. Detail Category

Following data elements MUST be present within the detail category.

   o Name of the region

   o Name of the RBridge

   o RBridge up time

   o Total number of neighbors

   o Total number of TRILL enabled ports in the RBridge

   o Total number of TRILL enabled ports Up

   o Total number of TRILL enabled ports oversubscribed

   o Total number of TRILL enabled ports observing errors

   o Maximum number of links in the largest ECMP of the switch

   o Port data: Name of each TRILL enabled Port and Port state (Up,
     oversubscribed, error) and interface index.

   o Adjacency Matrix

       o List of {neighbor RBridge nickname and interface index of
         ports connecting to the neighbor RBridge}.

       o NOTE: Interface index in the Adjacency matrix is used as
         key in to port data to obtain Port name and state.

```
+----------+
| subType  |              2 octets
+----------+
| RBridge  |
+----------+----------+
| Region-ID           | 4 octets
+--------------------+
| L|                  |
.--+                  .
.    Region Name      .
|                     |
+----------+----------+
|   UpTime            |
|                     | 8 octets
+--------------------+
| #Ports   |              2 octets
+----------+
| #Up Ports|              2 octets
+----------+
| #OsubPort|              2 octets
+----------+
| #ErrPorts|              2 octets
+----------+
| #ECMP    |              2 octets
+----------+
| #DwnPorts|              2 octets
+----------+
| subtype-2|              2 octets
+----------+----------+
|                     |
.   Port Data         .
.                     .
|                     |
+----------+----------+
| subtype-3|               2 octets
+----------+----------+
|                     |
.  Adjacency Matrix   .
.                     .
|                     |
+--------------------+
```
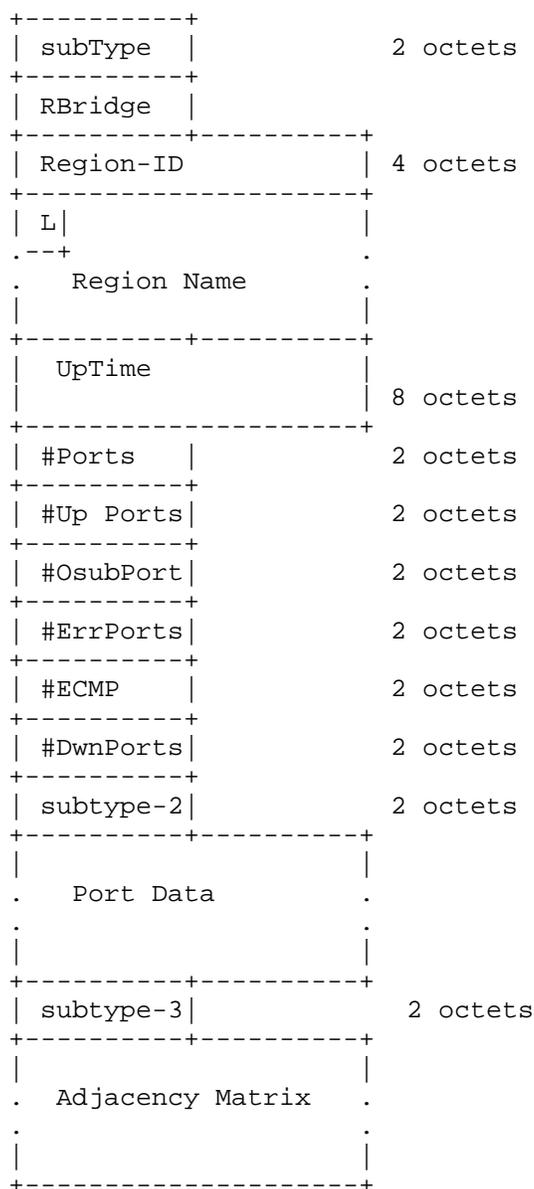
Figure 31   Encoding Detail Category Data

subType : (2 octets) always 2 for Detail category

RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

Regiond-ID : (4 octets) unsigned 32 bit integer identifier of the region

L   : (1 octet), length of the subsequent field

Region Name : '\0' terminated ASCII string of region name of variable size to maximum of 255 octets.

Up Time: (8 octets), number of seconds RBridge has been operational. If an RBridge reaches maximum count, it MUST NOT rollover.

#Ports: (2 octets) Total number of TRILL enabled ports available on this RBridge

#Up Ports: (2 octets) Total number of TRILL enabled ports that are operationally up.

#OSPorts : (2 octets) Total number of TRILL enabled ports that are oversubscribed.

#ErrPorts : (2 octets) Total number of TRILL enabled ports that are indicating errors.

#DwnPorts : (2 octets) Total number of TRILL enabled ports that are operationally down.

#ECMP : (2 octets) Maximum number of ECMP as seen by this RBridge ISIS routing table.

subtype-2: (2 octets): Set to 3. Following this sub type is the variable length Port Data. See below for details

sutype-3: (2 octets): Set to 4. Following this sub type is the variable length Adjacency Matrix. See below for details

```
     +----------+
     | subType  |              2 octets
     +----------+
     | RBridge  |
     +----------+----------+
     | F |                  1 octets
     +----------+
     | subtype-p|              2 octets
     +----------+----------+
     | ifindex             | 4 octets
     +----------+----------+
     | Slot     | Port     |
     .----------+----------+
     | Speed    | State    |
     +--------------------+
```
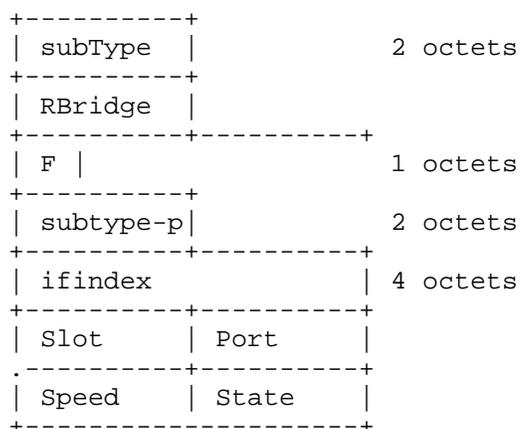
                     Figure 32   Encoding Port data

   subType : (2 octets) Set to3 for Port Data

   RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

   Regiond-ID : (4 octets) unsigned 32 bit integer identifier of the
   region

   L   : (1 octet), length of the subsequent field in octets.

   Region Name : '\0' terminated ASCII string of region name of
   variable size to maximum of 255 octets.

   F : (1 octet) Flag. When set, indicates this is the last Port data
   set from this node. It is possible Port data encoding to exceed MTU
   size due to large number of interfaces. The F flag allows to for
   advertising the information in multiple LSP packets.

   subtype-p: (2 octets) set to 5 to indicate that this is a single
   Port entry within subtype 3. SubType 5 MUST always be embedded with
   subtype 3. Within subtype 3 there can be multiple subtype 5, one for
   each port entry.

   Ifindex : (4 octets) 32 bit unsigned integer, used as key to port
   data advertised.

   Slot   (2 octets) : Slot number

   Port   (2 octets) : Port number

Speed   (2 octets) : Speed in 100Mbps. Zero (0) indicates port
speeds less than 100Mbps.

State   (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Link Monitoring disable
All other values reserved.

```
    +----------+
    | subType  |           2 octets
    +----------+
    | RBridge  |
    +----+-----+
    | F  |                 1 octets
    +----+-----+
    | subtype-a|           2 octets
    +----------+
    | nrBridge |           2 octets
    +----------+
    | #ports   |           2 octets
    +----------+----------+
    | ifindex            | 4 octets
    +--------------------+
    .                    .
    +--------------------+
```

Figure 33   Encoding Adjacency Matrix

subType : (2 octets) set to 4 for Adjacency Matrix

RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

Regiond-ID : (4 octets) unsigned 32 bit integer identifier of the
region

L   : (1 octet), length of the region name in octets

Region Name : '\0' terminated ASCII string of region name of
variable size to a maximum of 255 octets.

F : (1 octet) Flag. When set, indicates this is the last Port data
set from this node. It is possible Port data encoding to exceed MTU
size due to large number of interfaces. The F flag allows to for
advertising the information in multiple LSP packets.

subtype-a: (2 octets) set to 6 to indicates a single adjacency entry
within subtype 4. SubType 6 MUST always be embedded with subtype 4.
Within subtype 4, there can be multiple subtype 6, one for each
adjacency.

nrBRIDGE : (2 octets), nickname of the next hop RBridge

#ports : (2 octets), total number of parallel links from RBridge to
nrBRIDGE

Ifindex : (4 octets) 32 bit unsigned integer, used as key to port
data advertised.

10.5. Vendor Specific Category

Vendors may specify additional data elements to be distributed as
part of the monitoring data suite. All vendor specific data elements
MUST contain the regions name and follow the structure defined
below.

```
+----------+
| subType  |               2 octets
+----------+
| RBridge  |               2 octets
+----------+----------+
| Region-ID           |   4 octets
+---+-----------------+
| L |                 |
.---+                 .
.   Region Name       .
|                     |
+----------+----------+
| Vendor OUI          |   4 octets
+-------------------+
|                   |
. Vendor specific   .
.  Information      .
|                   |
+-------------------+
```
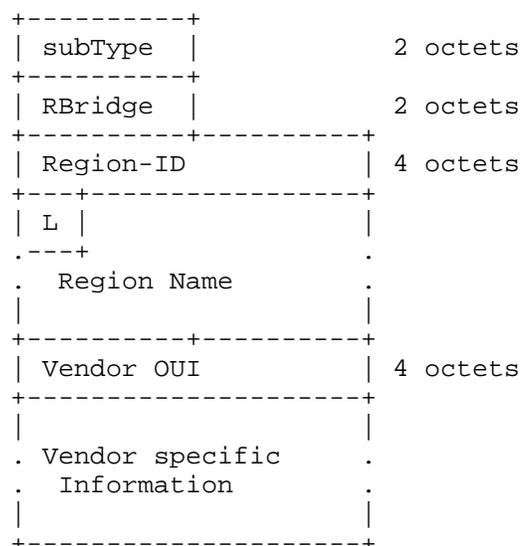
            Figure 34   Encoding Vendor specific category Data

    subType : (2 octets) set to250 for Vendor specific category

    RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

    Regiond-ID : (4 octets) unsigned 32 bit integer identifier of the
    region

    L   : (1 octet), length of the region name in octets

    Region Name : '\0' terminated ASCII string of region name of
    variable size to maximum of 255 octets.

    Vendor OUI : 3 octets of IEEE vendor OUI. Right justified. Most
    significant octet in network byte order is set to zero and ignored
    on recipt.

    Vendor specific information : variable size and vendor dependent.

11. Security Considerations

    Security considerations are under investigation.

12. IANA Considerations

12.1. IANA considerations

   Following IANA considerations are required

12.1.1. ICMP Extensions

   Request IANA to assign new Class-Num for TRILL OAM ICMP extensions.

   Request to form a sub-registry under ICMP extensions to include c-
   types defined in this document and allocate future requests.
   Currently c-types 1-20 are defined in section 8.1.

12.1.2. ARP Extensions

   Request IANA to form a new registry to allocate ARP extensions
   defined in section 9.5.1. . Class-Num allocated within ARP
   extensions are allocated by IANA on first come first serve basis. C-
   type within a given Class-Num are defined by owners of the Class-Num
   and sub-registry MUST be established within ARP extensions.

12.1.3. Well known Multicast MAC

   Request IETF authority to allocate one of the TRILL allocated
   Multicast MAC address (01-80-C2-00-00-43 to 01-80-C2-00-00-4F)for
   the purpose.

12.2. IEEE Registration Authority Consideration

   Well known unicast MAC address for the purpose of identifying OAM
   frames.

   Well known unciast MAC address for the purpose of identifying
   certain OAM frames.

   EthType <TBD> for the purpose of identifying OAM frames.

13. References

13.1. Normative References

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

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

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

   [RFC6327] Eastlake, Donald. et.al. "Routing Bridges (RBridges):
             Adjacency", RFC 6327, July 2011.

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

   [GenApp] Ginsberg, L. et.al. "Advertising Generic Information in
             IS-IS", draft-ietf-isis-genapp-04.txt, November,2010.

   [RFC4884] Bonica, R. et.al "Extended ICMP to support Multi-Part
             messages", April, 2007.

   [RFC4379] Kompella, K, and Swallow, G. "Detecting Multi-Protocol
             Label Switched (MPLS) Data Plane Failures", February,
             2006.

   [TRILLCH] Eastlake, Donald. et.al. "RBridges: TRILL RBridge Channel
             Support", draft-ietf-trill-channel-02.txt, July, 2011.

   [TRILLOAM]Bond, D. and Manral, V. "RBridges: Operations,
             Administration and Maintenance (OAM) Support", draft-ietf-
             trill-rbridge-oam-00.txt, July, 2011.

13.2. Informative References

   [RFC792] Postel, J. "Internet Control Message Proctocol (ICMP)",
             September, 1981.

   [RFC826] Plummer, D. "Address Resolution Protocol", November, 1982.

   [RFC2390] Bradley, T. et.al. "Inverse Address Resolution Protocol",
             September, 1988.

   [RFC5226] Narten, T. and Alverstand, H. "Guidelines for writing an
             IANA sections in RFCs", May 2008.

14. Acknowledgments

   Authors wish to thank people who volunteered to review this document
   and provided comments.

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

Appendix A.   Reports

A.1. Sample Reports

   In this section we present sample reports of summary data and sample
   output of detail data.

A.2. Summary Report

Region	Number of switches	Max ECMP	Total# Of Ports	% of Up Ports	%of Ports Oversubscribed	Err Ports
xxx	40	16	400	100	10	1
yyy	8	2	25	75	6	0

A.3. Detail Report

   Region Name : <xx>

   Total Number of Switches in the region   : 10
   Total Number of Core Ports in the region : 16
   Number of Operationally up Core Ports   : 14
   Number of Oversubscribed Core Ports   : 2
   Number of Error Core Ports   : 0

   Maximum Switch Up Time : 15days:8Hr:10M:0S
   Minimum Switch Up Time : 0days:0Hr:1M:0S

   Switch Adjacency Matrix:

   (*) oversubscribed Links
   (x) down Links
   (?) error Links

   Switch   Next Hop switch   Interfaces
    S1   S2   eth81,eth8/2(*),eth81
                                                  eth 10/2(x)

                    S3   eth5/1 (?)

                    S4   eth5/2,eth7/1

    S2   S1   eth4/1,eth4/2,eth3/1
                                                 eth3/2(x)

Authors' Addresses

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

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

Dinesh G Dutt
CISCO Systems
3800 Zankar Road
San Jose, CA 95134

Email: ddutt@cisco.com

Vishwas Manral
Hewlett-Packard Co.
19111 Pruneridge Ave.
Cupertino, CA 95014

Phone: 408-447-0000
EMail: vishwas.manral@hp.com

                 OAM tool for RBridges: Multi-destination Ping
                draft-yizhou-trill-multi-destination-ping-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/1id-abstracts.txt

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

   This Internet-Draft will expire on April 31, 2009.

Copyright Notice

carefully, as they describe your rights and restrictions with respect
to this document.

Abstract

Unicast and multi-destination data frame may follow the different
path in TRILL network. We need the ping and traceroute like
applications for the connectivity testing and fault isolation on the
multi-destination path in addition to the unicast path. This document
specifies the format and handling of the new TRILL OAM protocol
messages and TLVs which can be used for the multi-destination OAM.

Table of Contents

1. Introduction

   When RBridges are deployed in a real network, a number of
   applications are necessary for error detection/reporting and
   diagnostic purpose. TRILL RBridge channel [RBridgeChannel] was
   designed for carrying the OAM relevant messages. [RBridgeOAM] has
   defined the ping and traceroute applications for unicast path and
   also the error reporting mechanisms.

   Multi-destination data path in TRILL network has different
   characteristics from the unicast path. One or more distribution trees
   are formed for multi-destination traffic. RBridges advertise their
   interests in receiving the traffic of the specific VLANs. The
   distribution tree may or may not be pruned based on VLAN ID.
   Troubleshooting on the multi-destination path is a desirable feature
   of TRILL OAM. This document specifies the messages and mechanisms
   used by multi-destination OAM.

2. Conventions used in this document

   The same terminology and acronyms are used in this document as in
   [RF6325].

   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. Motivations

   In an RBridge campus, unicast and multi-destination traffic may
   follow different paths between the same ingress and egress RBridges.
   [RBridgeOAM] specifies some OAM along unicast path. For diagnostic
   purposes it is also desirable to check the connectivity between two
   or more RBridges along a particular distribution tree.

   There are various things we want to test for multi-destination path.

   - Along a distribution tree, who are the leaf nodes of an inner VLAN?
   The leaf nodes here refer to the RBridges announcing the given inner
   VLAN as their interested VLAN in INT-VLAN sub-TLV. It is useful when
   we want to check if the configuration/provisioning are consistent
   with the design.

   - Along a distribution tree, check the connectivity from the ingress
   RBridge to one or more leaf nodes of an inner vlan. It can be used as
   the first step in diagnosis when we suspect multi-destination data
   path to certain RBridge fails. Transit nodes do not decapsulate the

multi-destination data frame; therefore we do not think it is much of interest to check the connectivity to any non-leaf RBridges.

- Along a distribution tree, trace the multi-destination data path hop-by-hop to a target RBridge. It is useful when we want to find out where exactly is the failed hop.

This document specifies new messages and TLVs used by multi-destination OAM applications like multi-destination ping and traceroute. Processing of these messages is also discussed in the draft.

4. RBridge Channel Message Format

The RBridge Channel Header fields is as follows,

o CHV (Channel Header Version): zero.
o Channel Protocol: 0x006 (Echo in the Long Format) (TBD)
o Flags: The SL and NA bits SHOULD be zero, the MH bit SHOULD be one
o ERR: zero.

5. OAM Protocol Frame Format for Echo in the Long Format

The frame format is shown as follows. In the rest of this document, echo request and echo reply are brief ways to refer to the Echo Request in Long Format and Echo Reply in Long Format messages.

```
            | 0| 1| 2| 3| 4| 5| 6| 7| 8| 9|10|11|12|13|14|15|
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |                RBridge Channel                 |
            |                    Header                      |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |               Sender's Instance                |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |    SPID     |           Sequence               |
            |             |            Number                |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |    Reply mode          |         Flags         |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |                                                |
            |              TimeStamp Sent (48)               |
            |                                                |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |                                                |
            |             TimeStamp Received (48)            |
            |                                                |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            .                                                .
            .                    TLVs                        .
            .                                                .
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
                    Figure 1 Echo Request with Long Format
```

   o Sender's Instance: An instance ID used by sender to associate
the echo operation with different application instances, e.g.
different Telnet sessions. Echo reply should return the value
unchanged.

   o SPID:
          1 - Echo Request in the Long Format
          2 - Echo Reply in the Long Format

   o Sequence Number: An arbitrary 28-bit unsigned integer used to
aid in matching reply messages to echo requests. It MAY be zero.
   o Reply Mode: Default is 2. It can take one of the following
values.
          1 - Do not reply. It can be used for one-way connectivity
check. The receiving RBridge may perform monitoring and statistics
collection on delay and/or jitter using one-way echo operation.

        2 - Reply with Echo Reply in the Long Format and send back
   unicast in TRILL OAM channel. This value would be used by echo
   request in most cases.
     o Flags: A bit vector with the following format. Currently only
   the H (Respond Only When Hop Count is Zero) flag is defined. In
   practice, we set H flag to be 0 for ping type applications and 1 for
   traceroute type applications of multi-destination OAM. With H flag
   set, it will help to prevent the duplicate echo replies from the same
   RBridge triggered by echo request with different hop count value in
   the same traceroute operation. H flag is only significant in echo
   request and MUST NOT be set in echo reply. The detailed processing
   based on the value of H flag is explained in section 7.2.

```
            | 0| 1| 2| 3| 4| 5| 6| 7|
            +--+--+--+--+--+--+--+--+
            |     MBZ           | H|
            +--+--+--+--+--+--+--+--+
```
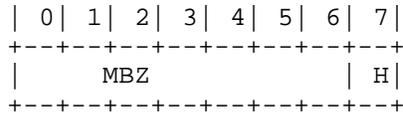
     o TimeStamp Sent: time-of-day (3 octets for seconds and 3 octets
   for microseconds) in NTP format that the echo request was sent
   according to the sender's clock.
     o TimeStamp Received: time-of-day (3 octets for seconds and 3
   octets for microseconds) in NTP format that the corresponding echo
   request was received according to the receiver's clock. This value is
   significant only in echo reply and MUST be set to all zeros in echo
   request and ignored on receipt of an echo request.
     o TLVs: A set of type, length, value encoded fields as specified
   in next section.

6. TLV Encodings

6.1. Target RBridge

```
            | 0| 1| 2| 3| 4| 5| 6| 7| 8| 9|10|11|12|13|14|15|
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |      Type = 0x05       |    Length = 2 + 2*n      |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            |            Number of Target RBridges          |
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
            .            Target RBridge Nickname 1          .
            .                     ...                       .
            .            Target RBridge Nickname n          .
            +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
```
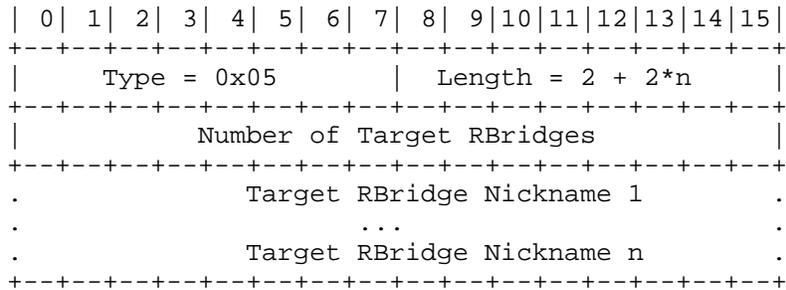
o Number of Target RBridges: The number of nicknames specified in the following fields, the maximum number is 127.

o Target RBridge Nickname: The nickname of a Target RBridge.

This TLV MAY appear in an echo request. It SHOULD be copied back in the corresponding echo reply messages.

Target RBridge TLV is used by multi-destination OAM. For ping along the multi-destination path, the Target RBridge TLV with multiple nicknames MAY be included in echo request. It implies RBridges with any of the nicknames in the TLV should reply. While for traceroute like application, only a single nickname can be included in this TLV. If there was more than one nickname in the TLV, only the first nickname MUST be used as target nickname for tracing purpose.

When Target RBridge TLV is not included in an echo request, it implies the unspecified target. If an echo request with unspecified target was sent by ping like applications, then all leaf nodes in distribution tree pruned by the given inner VLAN SHOULD send back echo reply. If echo request with unspecified target was sent by traceroute like application, RBridges receiving the incoming frame with hop count value 1 would process the echo request and send back 'Hop Count is Zero' error notification.

6.2. Jitter

```
        | 0| 1| 2| 3| 4| 5| 6| 7| 8| 9|10|11|12|13|14|15|
        +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
        |      Type = 0x07      |      Length = 0x02     |
        +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
        |                    Jitter time                |
        +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
```
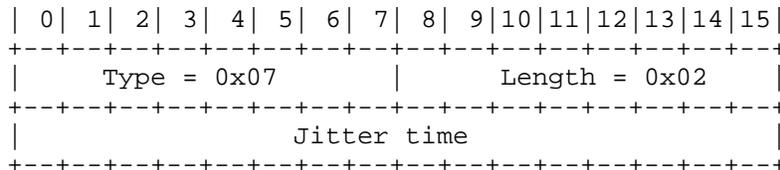
o Jitter time: Set to the upper bound of the jitter period in milliseconds. A responding node SHOULD wait a random amount of time between zero milliseconds and the value specified.

This TLV MAY appear in an Echo Request in the Long format. It SHOULD NOT be present in echo reply messages.

7. Processing Echo Messages for Multi-destination Path

7.1. Sending an echo request

   The inner frame header and TRILL header fields are as follows:

   o Inner.MacSA: MAC address of RBridge originating the echo request
   o Inner.MacDA: Defaults to All-Egress-RBridges. It can be set to L2
   multicast address derived from IP multicast group.
   o Inner.VLAN ID: Defaults to 1. It can be any enabled VLAN ID on the
   ingress RBridge.
   o Ingress RBridge Nickname: the nickname of RBridge originating the
     echo request
   o Egress RBridge Nickname: the nickname of a distribution tree root
   o M bit: 1
   o Hop Count: defaults to maximum value 0x3F.

      - For ping like applications, it can be any value which is believed
      to be no less than the number of hops from ingress RBridge to the
      most distant target RBridge in the tree.

      - For traceroute like applications, hop count value starts from 1
      and is increased by one for each sending of echo request.

   H(Respond Only When Hop Count is Zero) flag in echo request is set to
   1 for traceroute like applications and 0 for ping like applications.

   The originating RBridge chooses the values of Sender's Instance and
   Sequence Number for the echo request. Sequence number should be
   increased by 1 for each new subsequent echo request of the same
   Sender's Instance. The Timestamp Sent is set to the time-of-day in
   NTP format [NTP] according to the sender's clock. The Timestamp
   Received is set to zero.

   The originating RBridge MAY use Target RBridge TLV to specify the
   target. For ping like applications, multiple nicknames MAY be present
   in one such TLV if sender wants to ping multiple targets at one time.
   For traceroute like applications, the TLV should at most contain one
   nickname as the tracing target. If there is more than one nickname,
   only the first one takes effect.

Echo request without Target RBridge TLV means the originating RBridge potentially wants to target every RBridge in the distribution tree. We also call it echo request with the unspecified target. For ping like applications, echo request with the unspecified target implies the sender wants to know who are the leaf nodes of the inner VLAN in the distribution tree. For traceroute like applications, it implies the sender wants to know the whole distribution tree structure hop-by-hop.

The Originating RBridge MAY include the Jitter TLV (see section 6.2) in the echo request in order to randomize the delay of the replying echo message from multiple RBridges.

7.2. Receiving an echo request

RBridge receiving an echo request with M bit set with EtherType of RBridge channel [RBridgeChannel] SHOULD replicate it to the control plane for processing and also forward it as normal multi-destination data frame. When a RBridge receives an incoming frame with hop count is 1 in TRILL header, it will not forward the frame further. If reply mode is 1, no echo reply is generated. For the sub sections below, we assume the reply mode is set to 2.

7.2.1. If H (Respond Only When Hop Count is Zero) Flag Is Not Set

- If Target RBridge TLV is not present in the echo request:

   All leaf nodes of the distribution tree in the inner VLAN MUST process the incoming echo request and send back echo reply.

- If Target RBridge TLV is present in the echo request:

   An RBridge owning any one of the specified target nicknames in the incoming echo request MUST send back echo reply when it is a leaf node of the distribution tree in the inner VLAN.

If echo reply has already been generated for the incoming echo request, RBridge will not generate 'Hop Count is Zero' error notification even when the hop count value in the incoming echo request is one.

Echo request with 'H' flag unset is for ping like application. It should be noted that if an RBridge receives an echo request with its own nickname listed as one of the targets, it does not send back the echo reply if the RBridge did not advertise its interest of inner

VLAN. That is to say, the connectivity check using ping in multi-
destination path is constrained by inner VLAN. Normally VLAN 1 is the
default VLAN and enabled on every RBridge. Therefore it is
recommended to put inner VLAN to be 1 when we want to check the
connectivity without the constraint of a particular customer VLAN. We
may use the echo replies from that to plot the whole distribution
tree.

7.2.2. If H (Respond Only When Hop Count is Zero) Flag Is Set

When hop count of the incoming echo request is not one, RBridge would
never generate any echo reply or 'hop count iz zero' error
notification.

If the hop count is one in the incoming echo request:

- If Target RBridge TLV is not present in the echo request:

    RBridge receiving the incoming frame with hop count equal to 1
    MUST send back error notification of 'Hop Count is Zero'. RBridges
    MUST not generate any echo reply in this case. If hop count in
    incoming echo request is more than 1, control plane will not do
    anything. RBridge forwards the frame as normal multi-destination
    TRILL frame in data plane.

- If Target RBridge TLV is present in the echo request:

    RBridges owning the only target nickname listed in TLV MUST send
    back echo reply if it is a leaf node of the inner VLAN in the
    distribution tree. If it is not a leaf node of the inner vlan, no
    echo reply will be generated by the owner RBridge; however, 'Hop
    Count is Zero' error notification will be sent back instead.

    If an RBridge not owning the only target nickname listed in TLV
    receives the incoming frame with hop count equal to 1, it SHOULD
    check its LSDB. If it sits in-between of the ingress RBridge and
    the target RBridge along the specified distribution tree, RBridge
    MUST send back the error notification of 'Hop Count is Zero';
    otherwise the RBridge should not generate such error notification.
    The purpose of suppressing the error notification here is to make
    sure the ingress only receives the error notification along the
    real data path and to reduce the processing burden at ingress.

    If RBridge not owning the first target nickname listed in TLV
    receives the incoming frame with hop count greater than 1, the
    frame is forwarded as usual.

Echo request with 'H' flag set is for traceroute like applications. For traceroute with unspecified target, the ingress RBridge will be able to construct the whole distribution tree (when tree is not pruned) or the distribution tree of inner vlan (when tree is pruned by inner VLAN) according to the returned error notifications. For traceroute with a specified target in an inner VLAN, the ingress RBridge will receive the error notifications from the RBridges along the path to the target in the tree. If the target announced its interest of the inner VLAN, it will finally send back echo reply to the ingress. If the target did not announce its interest of the inner VLAN, either the target will not receive the echo request (e.g. it is located in the tree path being pruned) or the target will send back error notification of 'Hop Count is Zero' instead of echo reply.

7.3. Sending an echo reply

The inner frame header and TRILL header fields are as follows,

o Inner.MacSA: The MAC address of the RBridge generating the echo reply
o Inner.MacDA: All-Egress-RBridges
o Inner.VLAN ID: same as Inner.VLAN ID in the received echo request to which the echo reply responds
o Ingress RB Nickname: the nickname of the RBridge generating the echo reply.
o Egress RBridge Nickname: the ingress RBridge nickname in the corresponding received echo request
o M bit: 0
o Hop Count: defaults to the maximum value 0x3F. It can be any value that is believed to be larger than the number of hops from ingress to egress RBridge.

The values of Sender's Instance, Sequence Number and Timestamp sent in an echo reply MUST be same as those in its corresponding echo request. H flag MUST be zero in echo reply. The value of Timestamp Received is set to the time-of-day in NTP format [NTP] according to the receiver's clock.

If Target RBridge TLV was present in the echo request, the corresponding echo reply SHOULD copy it.

Next Hop Nickname and Incoming Port ID TLV [RBridgeOAM] MAY be included in echo reply.

When an echo reply is going to be sent to the originator RBridge, 'Hop Count is Zero' error notification MUST not be sent in response to the same echo request.

7.4. Receiving an echo reply

An RBridge SHOULD use the Sender's Instance and Sequence Number to match up the received echo reply with the echo request it sent. If there is no match found, the RBridge should discard the echo reply.

If Jitter TLV was present in the echo request, the round trip time should not be calculated based on the difference between the arriving time of echo reply and the value of "TimeStamp sent" in the replying frame. However the single trip time is always correct to be calculated on Timestamp Received minus Timestamp Sent when the clocks of sender and receiver are synchronized.

When an RBridge receives either an echo reply or 'hop count is zero' error notification from the target RBridge for traceroute like application, it SHOULD stop sending echo request with increased hop count value.

8. Security Considerations

The security vulnerabilities raised in [RBridgeOAM] also apply to the multi-destination RBridge ping in this document. The same mechanisms can be used to prevent or alleviate the security issues.

9. IANA Considerations

New error notification sub-code needs to be allocated by IANA as specified in Section 7.

10. References

10.1. Normative References

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

   [RBridgeChannel] Eastlake, D., Manral, V., Yizhou, L., Aldrin, S.,
         and D. Ward, "RBridges: TRILL RBridge Channel Support", draft-
         ietf-trill-rbridge-channel-02 (work in progress), July 2011.

   [RBridgeOAM] D. Bond, and V. Manral, "RBridges: Operations,
         Administration, and Maintenance (OAM) Support", draft-ietf-
         trill-rbridge-oam-01 (work in progress), October 2011.

   [NTP] Mills, D., "Simple Network Time Protocol (SNTP) Version 4 for
         IPv4, IPv6 and OSI", RFC 2030, October 1996.

10.2. Informative References

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

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

11. Acknowledgments

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

Authors' Addresses

   Yizhou Li
   Huawei Technologies
   101 Software Avenue,
   Nanjing 210012
   China

   Phone: +86-25-56624558
   Email: liyizhou@huawei.com

   Weiguo Hao
   Huawei Technologies
   101 Software Avenue,
   Nanjing 210012
   China

   Phone: +86-25-56623144
   Email: haoweiguo@huawei.com

David Michael Bond
International Business Machines
2051 Mission College Blvd.
Santa Clara, CA 95054
US

Phone: +1-603-339-7575
EMail: mokon@mokon.net
URI: http://mokon.net

Vishwas Manral
Hewlett Packard Co.
19111 Pruneridge Ave,
Cupertino, CA 95014 USA

Phone: +1-408-447-1497
EMail: vishwas.manral@hp.com

Network Working Group   L. Yong
Internet Draft   D. Eastlake
Intended status: Informational   S. Aldrin
                                                         Huawei
                                                      J. Hudson
                                                        Brocade
Expires: April 2012   October 23, 2011

     Transparent Interconnection of Lots of Links (TRILL) over an MPLS
                    PSN (Packet Switched Network)
                  draft-yong-trill-trill-o-mpls-00.txt

Status of this Memo

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

Abstract

This informational document describes ways to interconnect TRILL
RBridges over WAN connections by using MPLS Pseudo Wire (PW) or
Virtual Private LAN Service (VPLS) with existing TRILL and MPLS
standards. It also describes the combination of RBridge and MPLS to
provide a hierarchical scalable L2VPN.

Table of Contents

1. Introduction

   The IETF TRILL (Transparent Interconnection of Lots of Links)
   standard [RFC6325] [RFC6326] provides optimal pair-wise data frame
   forwarding without configuration in multi-hop networks with
   arbitrary topology, and support for multipathing of both unicast and
   multicast traffic. TRILL enables a new method to construct a campus
   or data center network. Devices that implement TRILL are called
   RBridges.

   This document describes the use cases of TRILL over an MPLS PW or
   VPLS, and introduces a new hierarchical L2VPN architecture that uses
   RBridges and IP/MPLS and documents the related configurations and
   references for the proper interworking.

   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].

   Acronyms used in this document include the following:

   AC - Attachment Circuit

   CE - Customer Edge

   IS-IS - Intermediate System to Intermediate System

   MPLS - Multi-Protocol Label Switching

   PE - Provider Edge

   PPP - Point to Point Protocol

   PW - Pseudo Wire

   RBridge - Routing Bridge

   TRILL - Transparent Interconnection of Lots of Links

   VPLS - Virtual Private LAN Service

   VSI - Virtual Service Instance

2. Use Cases

   RBridge campuses at different locations may interconnect by networks
   that are implemented with different technologies to form one RBridge
   campus. This section describes use cases assuming that IP/MPLS
   technology is available. From the MPLS network view, an RBridge
   device acts as a Customer Edge (CE) device and connects to PE via an
   attachment circuit (AC). RBridges [RFC6325] support both point-to-
   point links and multi-access links. Section 2.1 describes point-to-
   point link, i.e. TRILL over either Ethernet or PPP point-to-point
   link that is over an MPLS network. Section 2.2 describes TRILL over
   a bridged LAN that is implemented by MPLS/VPLS. Section 2.3
   introduces a new hierarchical L2VPN solution that uses the RBridges
   and MPLS tiered architecture.

2.1. Point-To-Point Interconnection

   Two RBridges are interconnected by either Ethernet or PPP link that
   is over a MPLS network. A Pseudo wire (PW) is configured between a

pair of PEs to provide part of the point-to-point link between two RBridges. Figure 1 illustrates this architecture. Each RBridge device connects to a PE via AC and acts as a CE device. MPLS PSN is bordered at the PEs. The TRILL link across the IP/MPLS PSN makes the left site and right site into one RBridge campus.

MPLS already supports many pseudo wire transport encapsulations. [RFC4446] Two types of TRILL links between RBridges have been standardized: Ethernet [RFC6325] and PPP [RFC6361]. PW encapsulations for these two interfaces are specified in [RFC4448] and [RFC4618], respectively. When an RBridge port connecting to AC is configured with a point-to-point Ethernet link type, two PEs can be configured as a PW with Ethernet encapsulation [RFC4448]. The PW between two PEs can be auto-configured [RFC4447] or manually configured; the two RBridges then appear directly interconnected with an Ethernet link. When an RBridge port connecting to AC is configured with the PPP link type, two PEs MUST be configured as a PW with PPP encapsulation.[RFC4618] After the PW is established between two PEs, the two RBridges then appear directly interconnected with a PPP link. The TRILL link is automatically configured over an Ethernet link or PPP link. The PW provides transparent transport between ACs.

Note: 1) For Ethernet link configuration, PE SHOULD use the Raw mode and non-service-delimiting, which provides a transparent transport. 2) The PPP link configuration will be more efficient than the Ethernet point-to-point configuration; it saves about 16 bytes per frame by replacing the TRILL Outer.MacDA, Outer.MacSA, Outer.VLAN, and outer Ethertype with a PPP code point. [RFC6361]

```
                     <----------TRILL Link---------->
          *-------*    <----------PW---------->    *-------*
          |       | AC +----+    +---+   +----+ AC |       |
          |  RB   +----| PE |----| P |---| PE |----+ RB    |
          |Site A |    +----+    +---+   +----+    |Site B |
          |       |    {         PSN           }  |       |
          *-------*                               *-------*
          {            One RBridge Campus                 }
```
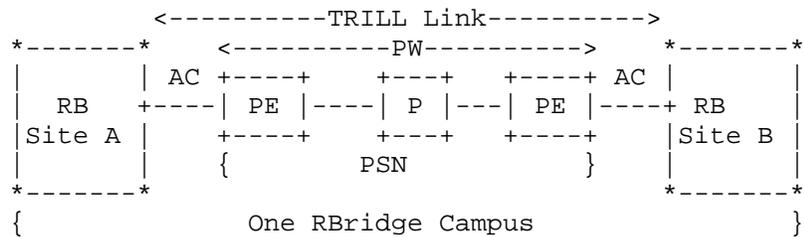
       Figure 1 P2P TRILL Link over IP/MPLS PSN Use Case I

An RBridge is a router and could support PE capabilities. As the networks converg, it is possible that one operator runs both an RBridge campus as well as the core MPLS network. Figure 2

illustrates this use case, in which RBridges are also MPLS PE
enabled devices. The interworking between the RBridge network and
the MPLS PSN is within the device. This has a similar architecture
to MPLS/VPLS [RFC4762]. In this case, a virtual Ethernet interface
is configured at the RBridge component; an Ethernet encapsulated PW
is configured between two interfaces, which brings up an TRILL link
between two RBridge components. The outer MAC address can be a local
Ethernet address. In this case, the Campus RBridges run in the
client layer and MPLS runs in the Server Layer; RB/PE devices
support both client and server layer control plane and data plane
functions.

```
      *---------*                          *---------*
      |         |<---- TRILL Link------->|         |
      |RB Site A|      (Client Layer)      |RB Site B|
      |    +-------+     +---+     +-------+    |
      |    | RB/PE |-----| P |------| PE/RB |    |
      |    +-------+     +---+     +-------+    |
      |         |<---------PW ---------->|         |
      |         |      (Server Layer)      |         |
      *---------*                          *---------*
                {           PSN           }
           {         One RBridge Campus           }
```
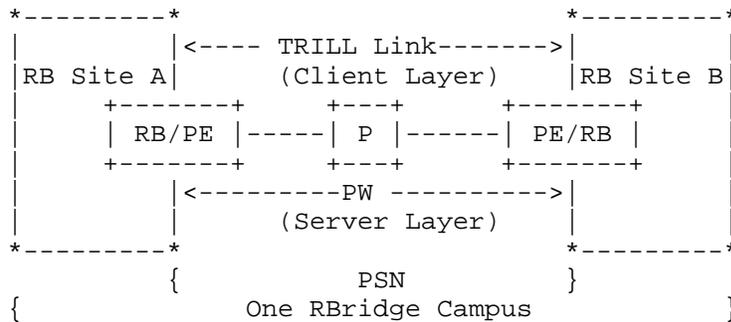
        Figure 2 P2P TRILL-Link over IP/MPLS PSN Use Case II

In both case I and II, the PE treats an RBridge as a generic CE and
has no awareness of TRILL capability on the CE. Use case I enables
the business models when the RBridge campus and Core MPLS may be
operated by different operators or the same operator. In the case of
different operators, the core MPLS operator can sell a VPWS service
to RBridge operator. Use case II provides the model when the RBridge
campus and the core network are operated by the same operator but
use different technologies in each network.

Technically speaking, it is possible to create a specially
designated TRILL encapsulated pseudo wire for point-to-point TRILL
over MPLS. However, the authors think that this is not worth the
effort because of available technologies as mentioned above,
particularly the highly-efficient PPP link technology.

A PW may cross multiple MPLS domains.[RFC5659] In this case,
RBridges connect to T-PEs and it works in the same way as a single
domain. The PSN can provide transport resiliency for a PW. The dual
homing (two ACs) can be used for AC protection. In this case, two

TRILL links are established; RBridge device perform load balance
over two links.

2.2. Multi-Access Link Interconnection

Multiple RBridges may interconnect via an 802.1Q Bridged LAN that
acts as a hub. The bridged LAN simply forwards on the outer Ethernet
header of the TRILL frames. This configuration creates what appears
to each connected RBridge as a multi-access link. In other words,
each RBridge connecting to a bridged LAN has connectivity to every
other RBridges connecting to the same LAN.

MPLS/VPLS can provide the same capability when multiple parts of an
RBridge campus are interconnected over an IP/MPLS PSN and make each
RBridge attaching to the VPLS to appear as having a multi-access
TRILL link. Figure 3 shows the use of MPLS/VPLS for RBridge
interconnection. One RBridge campus is split between three different
sites. One VPLS instance is configured on three PEs and the PWs are
configured for the VPLS instance. Each RBridge Site connects to the
VSI on a PE via an AC (Ethernet Type). The VSI on a PE forwards
TRILL frames based on the outer Ethernet header of the frames.
[RFC6325] Either BGP [RFC4761] or LDP [RFC4762] protocol can be used
to automatically construct the VPLS instance on the PEs. A PE may
connect to several different RBridge campuses that belong to
different customers. Separated VPLS instances are configured for
individual customers and customer traffic is completely isolated by
VPLS instance. The PE treats an RBridge as a generic CE and has no
awareness of TRILL.

```
           *-------*          .............................          *-------*
           |       |          .              MPLS/VPLS    .          |       |
           | RB    |AC +----+          PW           +----+AC| RB    |
           | Site 1+---|VSI ********************** VSI|--+ Site 2|
           |       |   | PE *****                *****  PE |  |       |
           |       |   +----+    ****     ****    +----+  |       |
           |       |          .          +*--*+         .          *-------*
           *-------*          ..........|VSI |...........
                                        | PE |
                                        +----+
                                         |AC
                                         |
                                *---------*
                                |         |
                                |RB Site 3|
                                |         |
                                *---------*
```
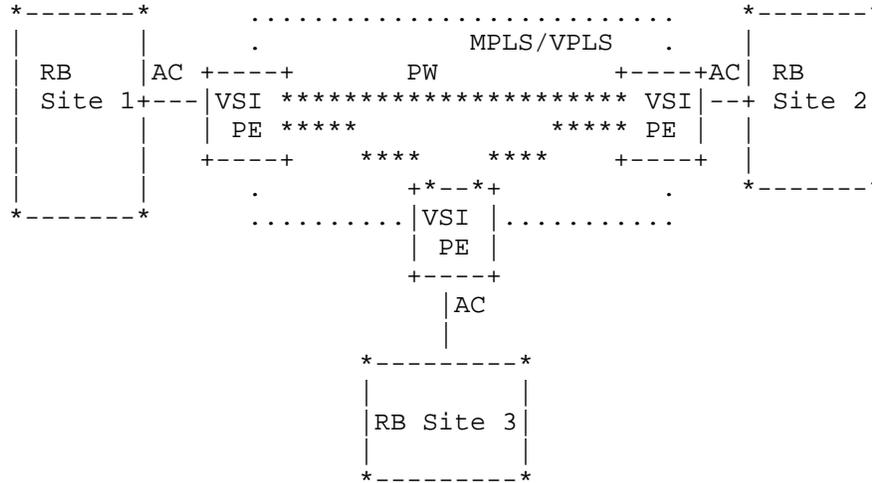
Figure 3 Multi-Access TRILL Link over MPLS/VPLS

The outer Ethernet MAC of TRILL frames may be either a next-hop
RBridge MAC address (for unicast frames) or one of TRILL defined
multicast addresses (ALL-IS-IS-RBridges and All-RBridges).[RFC6325]
The VSI at each PE learns the source MAC addresses on each VSI
interface and forward the frame based on the destination MAC. For
the multicast frames, the VSI replicates the frames to all PWs it
associates. If a VPLS is configured with some optimization
capability [VPLS-BCAST], the multicast frames can be delivered over
a point-to-multipoint PW while unicast frames are carried over a
point-to-point PW.

The scenario in Figure 3 can also be extended to multiple RBridges
interconnections when a device serves both RBridge and PE functions.
This use case is discussed in the following section.

Note: If the CEs associated with one VPLS instances happen to
include some RBridges and some end stations or IEEE 802.1Q bridges
to end stations, TRILL will, by default, be able to handle this by
providing both through service and end station service. However, the
end station addresses will be visible to the VPLS instance. If, in
such a case, all the RBridge ports connected to the VPLS are
configured as trunk ports (see Section 4.9.2 of [RFC6325]), then
they will not provide any end station service.

2.3. Hierarchical L2VPN with RBridges and MPLS

   H-VPLS in [RFC4762] describes a two-tier hierarchical solution for
   the purpose of pseudo wire (PW) scalability improvement. This
   improvement is achieved by reducing the number of PE devices
   connected in a full-mesh topology through connecting CE devices via
   the lower-tier access network, which in turn is connected to the
   top-tier core network. However, H-VPLS solutions in [RFC4762]
   require learning and forwarding based on customer MAC addresses,
   which poses scalability issues as the number of VPLS instances and
   customer MAC addresses increase. [PBB-VPLS] describes how to use PBB
   (Provider Backbone Bridges) at the lower-tier access network to
   solve the scalability issue, in which the transit network nodes only
   learn and forward on PBB port MAC addresses instead of customer MAC
   addresses.

   RBridges over IP/MPLS provide an alternative solution for a scalable
   L2VPN over WAN networks. Figure 4 depicts the hierarchical L2VPN
   architecture with RBridge/MPLS technologies. An IP/MPLS network
   serves as the top-tier core network function while an RBridge campus
   serves as the low-tier access network function. A RB/PE enabled
   device is placed at the boundary between the two-tier networks. A PW
   is configured between each pair of PE components in the top-tier
   IP/MPLS network, which constructs a full mesh TRILL links among the
   RB/PE devices. The RBridge component on a RB/PE device and other
   RBridges at the same site serves as the low-tier access network.
   Customer CEs connect to RBridges at each site directly. This
   architecture provides E-LAN or E-VLAN connectivity among customer
   CEs connecting to the RBridge campus sites. The transit RBridge node
   only forwards and learns other RBridge addresses and the number of
   PWs in the top-tier core network is not relate to the number of
   devices connecting to Customer CEs. This makes the solution scale
   very well. In addition, TRILL technology already supports multi-
   TRILL links from one RBridge to one or multiple RBridges and
   prevents loops, which provides the flexibility to construct the
   networks based on their network condition. Figure 4 shows that one
   RBridge in site 1 connects two RB/PE devices and one RB/PE device
   connects two RBridges at Site 2 via Ethernet links.

```
        +---------+      ...........................    +--------+
        |         |      .      IP/MPLS Core      .    |        |
        |         |      .                        .    |        +--
   CE--+ |Eth+----+      PW            +----+ |        |
        | RBridge +---| RB/|*******************| RB/|--+RBridge |CEs
   . --+         +-+ | PE |****           ****| PE |\ |        +--
   .    | Site 1 | | +----+    ****   ****   +----+ -+ Site 2 |
   . --+         | |    .     +*--*+          .    |        |
        |         | |    ..........| RB/|..........    +--------+
   CE--+         | +-------------| PE |
        |         | Eth. Itfc    +----+
        +---------+              |Eth. Itfc
                                 |
                        +---------+
                        | RBrdige |
                        |         |
                        | Site 3 |
                        +-+----+--+
                          |... |
                           CEs
```

                Figure 4 Hierarchical L2VPN with RBridge and MPLS

The following advantages of using RBridge/MPLS based L2VPN: 1)
Scalability improvement; 2) Auto-configuration; 3) Good efficiency
and loop prevention; and 4) Multipath support.

The solution also has the following advantages: 1) Since RBridge
terminates customer spanning tree protocol (STP), individual STPs in
attached customer bridged LANs will be separated and will converge
faster. 2) low over head per frame in number of added bytes and
scalable routing computations; 3) MPLS just provides P2P PWs, MAC
forwarding and learning does not exist within the MPLS network, thus
multi-homing issue does not exist.

Note: it is good to mention another scenario when the device has
both RB/PE capabilities, i.e. configure a VPLS instance among PE
components in the top-tier network to provide a multi-access link to
RBridge component on the RB/PE devices. Although this solution can
also provide scalability, it requires both the RBridge component and
VSI/PE component on a device to perform the same MAC forwarding and
learning functions, which is redundant. The number of PWs configured
in this case is the same as of the number of PWs in Figure 4. Thus,
authors do not recommend this configuration. For the same reason,
the use case in Section 2.2 is not viewed as the recommended L2VPN
solution for the WAN networks. Instead, it is useful when a Core

Service Provider provides a VPLS service to the customer who needs
to interconnect the RBridge campus sites over IP/MPLS PSN.

It is possible to construct a Tiered L2VPN in the combination of
Figure 4 and 3, i.e. some locations use RB/PE enabled device and
some location use separated RBridge and PE devices in a Hierarchical
L2VPN. When using separated RBridge and PE devices at some
locations, the MPLS network has to run a VPLS instance, which makes
RB/PE devices perform MAC forwarding and learning function two
times. In addition, it becomes operator responsibility to ensure
that the top tiered MPLS core is fully surrounded by an RBridge
campus. Missing configuration may increase the scalability problem
in the core network.

Auto configuration for the Hierarchical L2VPN will be addressed in
another draft.

3. RBridge Behavior for MPLS Pseudo Wire

This section describes RBridge behaviors for TRILL Ethernet or TRILL
PPP links over MPLS pseudo wire (PW) as described in Sections 2.1 .

1. For two RBridge ports connecting via a PPP PW, the ports MUST be
   configured as IS-IS point-to-point. Thus TRILL will use IS-IS P2P
   Hellos that, per "Point-to-Point IS to IS Hello PDU" (section 9.7
   of [IS-IS]), do not use Neighbor TLVs in the same manner as on a
   multi-access link. However, per section 4.2.4.1 of [RFC6325],
   three-way IS-IS handshake using extended circuit IDs is required.

2. For two RBridge ports connecting via an Ethernet PW, it is
   RECOMMENED that the ports be configured as IS-IS point-to-point
   for the same reason able. Note: an RBridge port by default
   supports multi-access links.

3. Any MPLS forwarder within an MPLS PSN does not change the TRILL
   Header Hop Count. RBridges is never aware of the packet
   forwarders in MPLS PSN.

4. If it is desired for MPLS PSN to perform QoS in the same way as
   in the RBridge campus, RBridges MUST be configured to send an
   Outer.VLAN tag on the RBridge port. The PE can then copy the
   priority value from the Outer.VLAN tag to the COS filed of the PW
   label prior to the forwarding. [RFC5462]

5. TRILL MTU-probe and TRILL MTU-ack messages (section 4.3.2 of
   [RFC6325]) are not needed on a pseudo wire link. Implementations
   MUST NOT send MTU-probe and SHOULD NOT reply to these messages.
   The MTU pseudo wire interface parameter SHOULD be used instead.
   PE Must configure the MTU size as the originating RBridges Size
   specified in Section 4.3.1 of [RFC6325].

4. Security Considerations

   The IS-IS authentication mechanism [RFC5304] [RFC5310], at the TRILL
   IS-IS layer, can be used to prevent fabrication of link-state
   control messages over TRILL links including those discussed in this
   document.

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

   The use case does not introduce any security considerations for MPLS
   network.

5. IANA Considerations

   No IANA action is required by this document.

6. Acknowledgements

   The authors sincerely acknowledge the contributions of Ben Mack-
   Crane and Sue Hares.

7. References

7.1. Normative References

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

   [RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge
             Emulation (PWE3)", BCP 116, RFC 4446, April 2006.

   [RFC4447] Martini, L., etc, "Pseudowire Setup and Maintenance Using
             the Label Distribution Protocol (LDP)", RFC4447, April
             2006.

   [RFC4448] Martini, L., "Encapsulation Methods for Transport of
             Ethernet over MPLS Networks", BCP 116, RFC 4446, April
             2006.

   [RFC4618] Martini, L., "Encapsulation Methods for Transport of
             PPP/High-Level Data Link Control (HDLC) over MPLS
             Networks", BCP 116, RFC 4618, September 2006.

   [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service
             (VPLS) Using BGP for Auto-Discovery and Signaling", RFC
             4761, January 2007.

   [RFC4762] Lasserre, M. and Kompella, V, "Virtual Private LAN Service
             (VPLS) Using Label Distribution Protocol (LDP) Signaling",
             RFC4762, January 2007

   [RFC5304] Li, T. and Atkinson, R, "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

   [RFC5462] Andersson, L. and Asati, R., "Multiprotocol Label
             Switching (MPLS)Label Stack entry: "Exp" Field Rename to
             "Traffic Class" Field", RFC5462, February 2009

   [RFC5659] Bocci, M and Bryant, S, "An Architecture for Multi-Segment
             Pseudowire Emulation Edge-to-Edge", RFC 5659, October
             2009.

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

   [RFC6326] Eastlake 3rd, D., Banerjee, A., Dutt, D., Perlman, R., and
             Ghanwani, A. "Transparent Interconnection of Lots of Links
             (TRILL) Use of IS-IS", RFC6326, July 2011.

   [RFC6361] Carlson, J., and D. Eastlake, "PPP Transparent
             Interconnection of Lots of Links (TRILL) Protocol Control
             Protocol", RFC6361, August 2011.

7.2. Informative References

   [IS-IS]   International Organization for Standardization,
             "Intermediate system to Intermediate system intra-domain
             routing information exchange protocol for use in
             conjunction with the protocol for providing the
             connectionless-mode Network Service (ISO 8473)", ISO/IEC
             10589:2002, Second Edition, Nov 2002

   [VPLS-BCAST] Delord, S, and Key, R., "Extension to LDP-VPLS for
             Ethernet Broadcast and Multicast", draft-ietf-l2vpn-ldp-
             vpls-broadcast-exten-02, work in progress, 2011.

   [PBB-VPLS] Sajarssi, A, etc, "VPLS Interoperability with Provider
             Backbone Bridges", draft-ietf-l2vpn-pbb-vpls-interop, work
             in progress, 2011

Authors' Addresses

   Lucy Yong
   Huawei Technologies (USA)
   5340 Legacy Drive
   Plano, TX 75025

   Phone: +1-469-277-5837
   Email: lucy.yong@huawei.com

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

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

   Sam Aldrin
   Huawei Technologies
   2330 Centeral Expressway
   Santa Clara, CA 95050

   Phone: +1-408-330-4517
   Email: sam.aldrin@huawei.com

   Jon Hudson
   Brocade
   130 Holger Way
   San Jose, CA 95134

   Phone: +1-408-333-4062
   jon.hudson@brocade.com

                        TRILL IS-IS MTU Negotiation
                 draft-zhang-trill-mtu-negotiation-01.txt

Abstract

   The IETF TRILL protocol provides least cost pair-wise layer 2 data
   forwarding by using IS-IS link state routing. This document defines a
   new link MTU size negotiation mechanism to update the TRILL documents
   "Routing Bridges (RBridges): Base Protocol Specification" and
   "Routing Bridges (RBridges): Adjacency".

Status of this Memo

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

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

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

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/1id-abstracts.html

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

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

   The base TRILL protocol includes the way how RBridges determine the
   minimum inter-RBridge link size for the whole campus (campus-wide
   Sz), for the proper operation of TRILL IS-IS. According to [RFC6325],
   RBridges need to know the campus-wide Sz before they do the link MTU
   size testing. The link MTU size testing therefore depends on the
   campus-wide Sz collection.

   [RFC6327] defines the diagram of state transitions of an adjacency.
   The "link MTU size is successfully tested (A6)" is an articulate
   transition between "2-way" state and "Report" state of an adjacency.
   It is not clear, in this draft, when an adjacency should start to
   synchronize LSP database.

   This document analyzes the possible issues caused by the definition
   that link MTU size testing depends on campus-wide Sz collection. A
   new link MTU size negotiation mechanism is provided to solve the
   above problems.

1.1. Content

   Section 2 analyzes the issues caused by the dependence on campus-wide
   Sz for link MTU size testing.

   Section 3 defines a new IS-IS MTU negotiation mechanism to update
   [RFC6325].

   Section 4 provides a method for link traffic MTU determination.

1.2. Terminology

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

2. Issues of Link MTU Testing

   Link MTU size testing is defined in Section 4.3.2 of [RFC6325]. If
   the link MTU size is smaller than campus-wide value of Sz, which is
   the smallest value of Sz advertised by any RBridge in its LSP
   [RFC6325], the link is not included in the global topology. If the
   link MTU size X of an adjacency is successfully tested (X >= campus-
   wide Sz), its state will move from 2-way to Report, which is defined
   in [RFC6327]. The link MTU size testing depends on the value of
   campus-wide Sz, which can be problematic. The issues causes by this
   dependence are given in the following subsections.

2.1. Global Dependence

```
     Sz:1800              Sz:1800                                Sz:1800
      +---+      +---+                     +--+            +---+
      |RB1|(2000)-(2000)|RB2| (2000)-(1700)|B1|(1700)-(2000)|RB3|
      +---+      ^      +---+                     +--+            +---+
      (2000)     |                                 ^
        |<---- Report                              |
      (2000)                                     Report
      +---+
      |RB4|
      +---+                                         |
     Sz:1600                                        v


     Sz:1800              Sz:1800                                Sz:1800
      +---+      +---+                     +--+            +---+
      |RB1|(2000)-(2000)|RB2| (2000)-(1700)|B1|(1700)-(2000)|RB3|
      +---+      ^      +---+                     +--+            +---+
                 |                                 ^
               Report                              |
                                                2-way
```
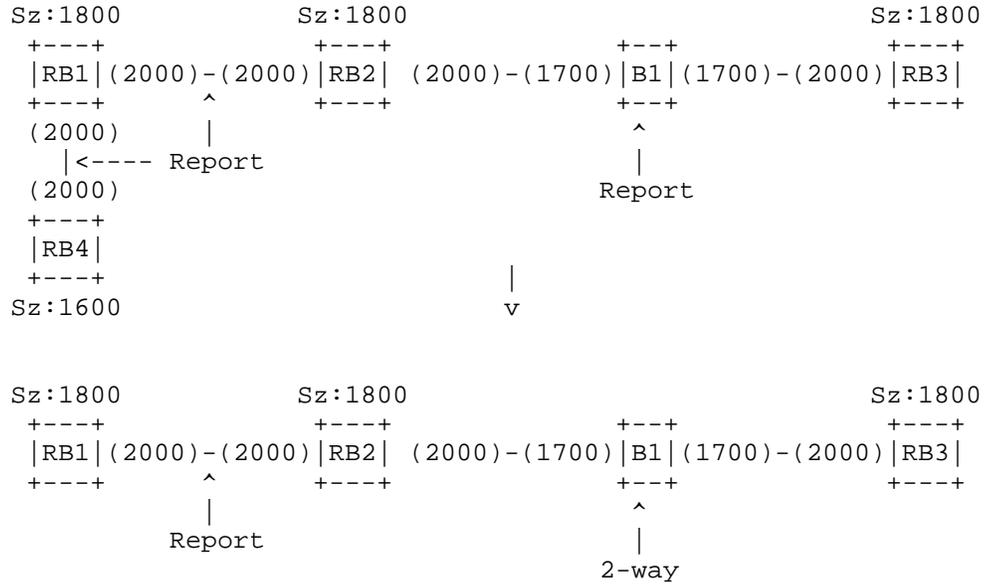
Figure 2.1: Adjacency global dependence

   Take Figure 2.1 as an example, all the adjacencies are in report
   states. After RB4 leaves the campus, RB2 and RB3 find the campus-wide
   Sz grows. They test the MTU according to campus-wide Sz 1800. Since
   RB2 and RB3 is connected by a low-end bridge whose port MTU is 1700.
   The test will not be successful. This adjacency has to return to 2-
   way state. The state of an adjacency can be determined by another
   remote adjacency. The stability of the campus Sz can be terrible
   resulting in maintenance problems.

2.2. Concealing Wrong Configuration

   Take Figure 2.2 as an example, the Sz value of RB3 is falsely
   configured to be greater than its port MTU. The link MTU testing is
   successful because the campus-wide Sz 1600 is smaller than the two
   port MTUs of the adjacency between RB2 and RB3. The adjacency will be
   in "Report" state. However, when RB4 leaves the campus and the
   campus-wide Sz is updated to 1800, the link MTU test of link RB2-RB3
   cannot be successful.

```
    Sz:1600            Sz:1800             Sz:1800             Sz:1800
     +---+              +---+               +---+               +---+
     |RB4|(2000)-(2000)|RB1|(2000)-(2000)|RB2|(2000)-(1700)|RB3|
     +---+      ^       +---+      ^        +---+      ^        +---+
               |                  |                   |
             Report             Report              Report
                                   |
                                   v

              Sz:1800             Sz:1800             Sz:1800
               +---+               +---+               +---+
               |RB1|(2000)-(2000)|RB2|(2000)-(1700)|RB3|
               +---+      ^        +---+      ^        +---+
                         |                   |
                       Report              2-way
```
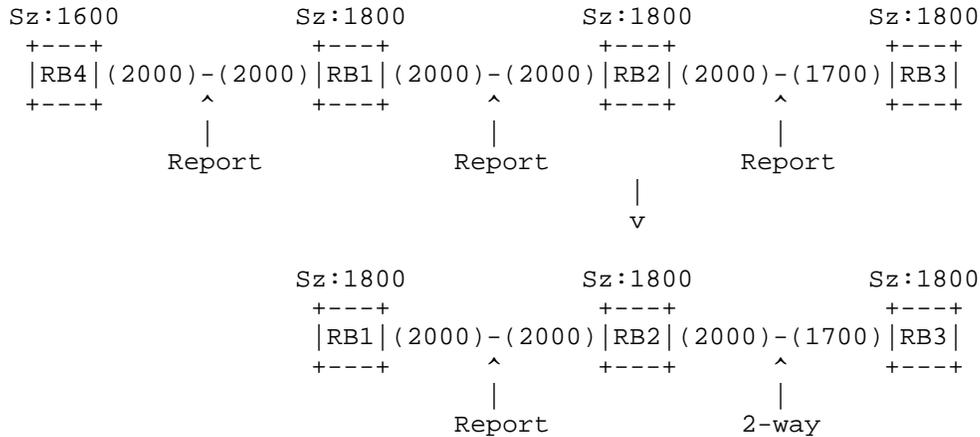
               Figure 2.2: Concealing wrong configuration

3. TRILL IS-IS MTU Negotiation

   It is improper to use campus-wide Sz in link MTU testing and LSP
   database synchronization. In order to solved the problems depicted in
   Section 2, this draft introduces a new value "Lz" which is the
   minimum acceptable inter-RBridge link size required by RBridges on a
   specific LAN link. Lz is used in link MTU size testing and LSP
   database synchronization to replace the role of campus-wide Sz. After
   link MTU size is successfully tested, the adjacency is changed to
   "Report" state.

3.1. Determination of Lz

   RBridges on a LAN link should exchange their local Sz through LSPs
   using the originatingLSPBufferSize, TLV #14. The smallest value of
   these Sz is Lz. Therefore, Lz is actually a "link-wide Sz". It is
   different from the campus-wide Sz which is determined by having each
   RBridge in the campus advertise its own assumption of the value of Sz
   in LSPs as defined in Section 4.3.1 of [RFC6325].

   The maximum size of some types of PDUs should be confined by Lz
   rather than campus-wide Sz because they are only exchanged between
   neighbors instead of the whole campus. CSNPs and PSNPs are such kind
   of PDUs. They are exchanged just on the link after a DRB is selected
   on the link.

```
                      Sz:1800            Sz:1800
                       +---+      |       +---+
                       |RB1|(2000)-|-(2000)|RB2|
                       +---+      |       +---+
                                  |
           Sz:1800                |
            +---+                +--+
            |RB3|(2000)-(1700)|B1|
            +---+                +--+
                                  |
```

               Figure 3.1: Link MTU has to be negotiated

    Even all RBridges on a specific LAN link have reached consensus on
    the value of Lz, it does not mean that these RBridges can safely
    exchange PDUs between each other. Take Figure 3.1 as an example. RB1,
    RB2 and RB3 are three RBridges on the same LAN link and their Sz are
    1800, so the link-wide Sz of this LAN link is 1800. There is a bridge
    (say B1) between RB2 and RB3 whose port MTU size is 1700. If RB2
    sends PDUs formatted in the size of 1800, it will be discarded by B1.
    Therefore the link MTU size has to be tested. Only after the link MTU
    size of an adjacency is successfully tested, these CSNP and PSNP PDUs
    will be formatted no greater than the tested link MTU size and will
    be safely transmitted on this link.

3.3. Link MTU Size Testing Algorithm

    The link MTU size testing method given by the last paragraph of
    Section 4.3.2 of [RFC6325] is updated by the following Binary Search
    algorithm in which Lz is used in the testing instead of campus-wide
    Sz.

    Step 0: RB1 sends an MTU-probe padded to the size of Lz.

    1) If RB1 successfully receives the MTU-ACK to the probe of size Lz
       from RB2, then link MTU size is set to the size of Lz and stop.

    2) RB1 tries to send an MTU-probe padded to the size 1470.

       a) If RB1 fails to receive an MTU-ACK from RB2 after k tries
          (where k is a configurable parameter whose default is 3), RB1
          sets the "failed minimum MTU test" flag for RB2 in RB1's Hello
          and stop.

       b) Link MTU size <-- 1470, X1 <-- 1470, X2 <-- Lz, X <-- [(X1 +
          X2)/2] (Operation "[...]" returns the fraction-rounded-up
          integer.). Repeat Step 1.

Step 1: RB1 tries to send an MTU-probe padded to the size X.

1) If RB1 fails to receive an MTU-ACK from RB2 after k tries, then:

   X2 <-- X and X <-- [(X1 + X2)/2]

2) If RB1 receives an MTU-ACK to a probe of size X from RB2 then:

   link MTU size <-- X, X1 <-- X and X <-- [(X1 + X2)/2]

3) If X1 >= X2 or Step 1 has been repeated n times (where n is a configurable parameter whose default is 5), stop. Else go to Step 1.

Since the execution of the above algorithm can be resource consuming, it is recommended that the DRB takes the responsibility to do the testing. If the testing is finished and the tested link MTU size is smaller than the original Lz and the minimum Sz that has been advertised to the DRB, the DRB should send the tested link MTU size as its local originatingLSPBufferSize in LSP number zero (shorted as LSP0). This will trigger other RBridges on the link to update their Lz to be the size of the tested link MTU. Then CSNPs, PSNPs and LSPs used for synchronization can be rightly resized and successfully exchanged on the link.

3.4. Re-determining Campus-Wide Sz

RBridges may join in or leave the campus from time to time. The campus-wide Sz can become outdated. Section 4.3.1 of [RFC6325] does not define when to re-determine the campus-wide Sz. The following suggestions are given for campus-wide Sz re-determination.

1) When a new RB whose Sz is smaller than current campus-wide Sz joins in the campus, it MUST report its Sz in an LSP which will cause other RBridges update their campus-wide Sz. The LSPs in the campus will be resized to be no greater than the new campus-wide Sz.

2) When an RB whose Sz is right the campus-wide Sz leaves the campus, and the LSPs generated by this RBridge are purged from the remaining campus after reaching MaxAge [ISO10589]. The campus-wide Sz ought to be resized as well. Frequent LSP "resizing" is harmful to the stability of the whole campus, so it should be dampened. Within the two kinds of resizing actions, only the upward resizing will be dampened. When an upward resizing event happens, a timer is set (this is a configurable parameter whose default value is 300 seconds). Before this timer expires, all subsequent upward resizing will be dampened.

   3) An RBridge may generate multiple LSPs. It is recommended that each
      RBridge carries its Sz in LSP0 [ISO10589]. Otherwise, if Sz is
      absent in LSP0, the campus-wide Sz will be set to a small value
      1470 at the receiver RBridge [RFC6325]. When subsequent LSPs
      carrying Sz arrives, the campus-wide Sz will be resized again.

3.5. Relationship between Port MTU and Sz

   When port MTU size is smaller than the local Sz of an RBridge, this
   port should be explicitly disabled from the TRILL campus. On the
   other hand, when an RBridge receives an LSP with size greater than
   its local Sz or the campus-wide Sz, this LSP should be normally
   processed rather than discarded. If an LSP is larger than the MTU
   size of a port over which it is to be propagated, no attempt shall be
   made to propagate this LSP over the port and an
   LSPTooLargeToPropagate alarm shall be generated [ISO10589].

3.6. LSP Synchronization

   The DRB of a LAN link is elected as early as in the "Detect" state of
   an adjacency. When a DRB is elected, it begins to send out CSNP to
   synchronize the LSP database of the RBridges attached to this LAN
   link when the adjacency between this RBridge and the DRB moves to 2-
   way state. If a non-DRB RBridge receives this CSNP and finds that
   LSPx is not in its LSP database, it will send out PSNP to request
   LSPx from the DRB. If a non-DRB receives this CSNP and finds that
   LSPx is not in the LSP database of the DRB, it will also send out
   LSPx to the DRB.

   DRB and non-DRB on a link should start to synchronize LSP database
   using CSNPs and PSNPs with a neighbor when the adjacency between them
   moves to the 2-way state [RBclr]. The CSNPs and PSNPs should be
   formatted in chunks of size at most Lz. Since the link MTU size has
   not been tested, Lz may be greater than the actually the link MTU
   size. In that case, an CSNP or PSNP may be discarded if its size is
   greater than the link MTU size. After the link MTU size is
   successfully tested, the adjacencies will begin to formatted these
   PDUs in the size no greater than it, therefore these LSPs will
   successfully get through.

4. Determining Link Traffic MTU Size

   Campus-wide Sz is used to confine the size of the TRILL link state
   information messages (LSPs). This value is different from the MTU
   size that restricting the size of TRILL data frames. TRILL data frame
   forwarded by an RBridge can be greater than the campus-wide Sz or Lz.
   They are restricted by the physical links and devices.

The algorithm defined in link MTU size testing can also be used in TRILL traffic MTU size testing, only that Lz used in that algorithm should be replaced with the port MTU of the RBridge sending MTU probes. The successfully tested size X can be advertised as an attribute of this link using MTU sub-TLV defined in section 2.4 of [RBisis]. An end station may collect these values by TRILL ping or traceroute. Path MTU is the smallest tested link MTU on this path.

5. Security Considerations

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

6. IANA Considerations

   No new registry is requested to be assigned by IANA.

7. References

7.1. Normative References

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

   [RBaf]   R. Perlman, D. Eastlake, et al, "RBridges: Appointed
              Forwarders", draft-ietf-trill-rbridge-af-05.txt, working
              in progress.

   [RFC6327] D. Eastlake, R. Perlman, et al, "Routing Bridges
              (RBridges): Adjacency", RFC 6327, July 2011.

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

   [RBclr]   D. Eastlake, M. Zhang, et al, "RBridges: Clarifications
              and Corrections", draft-eastlake-trill-rbridge-clear-
              correct-00.txt, working in progress.

7.2. Informative References

   [ISO10589] ISO, "Intermediate system to Intermediate system routeing
              information exchange protocol for use in conjunction with
              the Protocol for providing the Connectionless-mode Network
              Service (ISO 8473)," ISO/IEC 10589:2002.

Author's Addresses

    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

    Xudong 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: zhangxudong@huawei.com

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

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