INTERNET-DRAFT   M. Zhang
Intended Status: Proposed Standard   D. Eastlake
                                                               Huawei
                                                           R. Perlman
                                                                  EMC
                                                            M. Cullen
                                                    Painless Security
                                                              H. Zhai
                                                                  JIT
Expires: August 19, 2016   February 16, 2016

      Transparent Interconnection of Lots of Links (TRILL)
       Single Area Border RBridge Nickname for Multilevel
        draft-ietf-trill-multilevel-single-nickname-01.txt

Abstract

   A major issue in multilevel TRILL is how to manage RBridge nicknames.
   In this document, the area border RBridge uses a single nickname in
   both Level 1 and Level 2. RBridges in Level 2 must obtain unique
   nicknames but RBridges in different Level 1 areas may have the same
   nicknames.

Status of This Memo

Copyright and License Notice

Table of Contents

[1](#). Introduction

TRILL multilevel techniques are designed to improve TRILL scalability
issues. As described in [[MultiL](#)], there have been two proposed
approaches. One approach, which is referred as the "unique nickname"
approach, gives unique nicknames to all the TRILL switches in the
multilevel campus, either by having the Level-1/Level-2 border TRILL
switches advertise which nicknames are not available for assignment
in the area, or by partitioning the 16-bit nickname into an "area"
field and a "nickname inside the area" field. The other approach,
which is referred as the "aggregated nickname" approach, involves
assigning nicknames to the areas, and allowing nicknames to be reused

in different areas, by having the border TRILL switches rewrite the
nickname fields when entering or leaving an area.

The approach specified in this document is different from both
"unique nickname" and "aggregated nickname" approach. In this
document, the nickname of an area border RBridge is used in both
Level 1 (L1) and Level 2 (L2). No additional nicknames are assigned
to the L1 areas. Each L1 area is denoted by the group of all
nicknames of those border RBridges of the area. For this approach,
nicknames in L2 MUST be unique but nicknames inside different L1
areas MAY be reused. The use of the approach specified in this
document in one L1 area does not prohibit the use of other approaches
in other L1 areas in the same TRILL campus.

2. Acronyms and Terminology

Data Label: VLAN or FGL Fine-Grained Label (FGL)

IS-IS: Intermediate System to Intermediate System [IS-IS]

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].

Familiarity with [RFC6325] is assumed in this document.

3. Nickname Handling on Border RBridges

This section provides an illustrative example and description of the
border learning border RBridge nicknames.

```
        Area {2,20}              level 2            Area {3,30}
+-------------------+    +-----------------+    +--------------+
|                   |    |                 |    |              |
| S--RB27---Rx--Rz----RB2---Rb---Rc--Rd---Re--RB3---Rk--RB44---D |
|     27            |    |                 |    |      44      |
|             ----RB20---              ----RB30---            |
+-----------------+    +----------------+    +-------------+
```

        Figure 1: An Example Topology for TRILL Multilevel
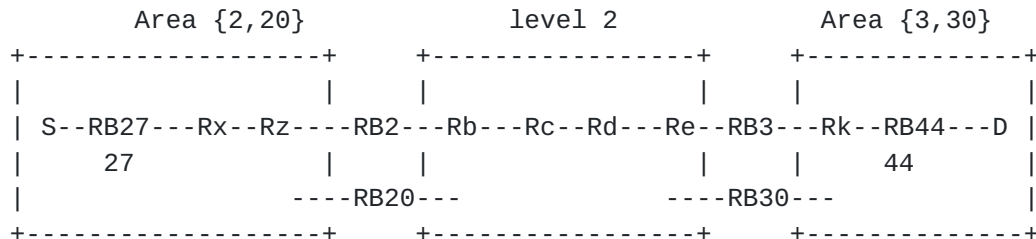
In Figure 1, RB2, RB20, RB3 and RB30 are area border TRILL switches
(RBridges). Their nicknames are 2, 20, 3 and 30 respectively. Area
border RBridges use the set of border nicknames to denote the L1 area
that they are attached to. For example, RB2 and RB20 use nicknames
{2,20} to denote the L1 area on the left.

A source S is attached to RB27 and a destination D is attached to

RB44. RB27 has a nickname, say 27, and RB44 has a nickname, say 44
(and in fact, they could even have the same nickname, since the TRILL
switch nickname will not be visible outside these Level 1 areas).

3.1. Actions on Unicast Packets

Let's say that S transmits a frame to destination D and let's say
that D's location is learned by the relevant TRILL switches already.
These relevant switches have learned the following:

1) RB27 has learned that D is connected to nickname 3.
2) RB3 has learned that D is attached to nickname 44.

The following sequence of events will occur:

- S transmits an Ethernet frame with source MAC = S and destination
  MAC = D.

- RB27 encapsulates with a TRILL header with ingress RBridge = 27,
  and egress RBridge = 3 producing a TRILL Data packet.

- RB2 and RB20 have announced in the Level 1 IS-IS instance in area
  {2,20}, that they are attached to all those area nicknames,
  including {3,30}. Therefore, IS-IS routes the packet to RB2 (or
  RB20, if RB20 on the least-cost route from RB27 to RB3).

- RB2, when transitioning the packet from Level 1 to Level 2,
  replaces the ingress TRILL switch nickname with its own nickname,
  so replaces 27 with 2. Within Level 2, the ingress RBridge field
  in the TRILL header will therefore be 2, and the egress RBridge
  field will be 3. (The egress nickname MAY be replaced with an area
  nickname selected from {3,30}. See Section 4 for the detail of the
  selection method. Here, suppose nickname 3 is used.) Also RB2
  learns that S is attached to nickname 27 in area {2,20} to
  accommodate return traffic. RB2 SHOULD synchronize with RB20 using
  ESADI protocol [RFC7357] that MAC = S is attached to nickname 27.

- The packet is forwarded through Level 2, to RB3, which has
  advertised, in Level 2, its L2 nickname as 3.

- RB3, when forwarding into area {3,30}, replaces the egress
  nickname in the TRILL header with RB44's nickname (44). (The
  ingress nickname MAY be replaced with an area nickname selected
  from {2,20}. See Section 4 for the detail of the selection method.
  Here, suppose nickname 2 is selected.) So, within the destination
  area, the ingress nickname will be 2 and the egress nickname will
  be 44.

  - RB44, when decapsulating, learns that S is attached to nickname 2,
     which is one of the area nicknames of the ingress.

3.2. Actions on Multi-Destination Packets

   Distribution trees for flooding of multi-destination packets are
   calculated separately within each L1 area and L2. When a multi-
   destination packet arrives at the border, it needs to be transitioned
   either from L1 to L2, or from L2 to L1. All border RBridges are
   eligible for Level transition. However, for each multi-destination
   packet, only one of them acts as the Designated Border RBridge (DBRB)
   to do the transition while other non-DBRBs MUST drop the received
   copies. All border RBridges of an area SHOULD agree on a pseudorandom
   algorithm and locally determine the DBRB as they do in the "Per-flow
   Load Balancing" section. It's also possible to implement a certain
   election protocol to elect the DBRB. However, such kind of
   implementations are out the scope of this document.

   As per [RFC6325], multi-destination packets can be classified into
   three types: unicast packet with unknown destination MAC address
   (unknown-unicast packet), multicast packet and broadcast packet. Now
   suppose that D's location has not been learned by RB27 or the frame
   received by RB27 is recognized as broadcast or multicast. What will
   happen, as it would in TRILL today, is that RB27 will forward the
   packet as multi-destination, setting its M bit to 1 and choosing an
   L1 tree, flooding the packet on the distribution tree, subject to
   possible pruning.

   When the copies of the multi-destination packet arrive at area border
   RBridges, non-DBRBs MUST drop the packet while the DBRB, say RB2,
   needs to do the Level transition for the multi-destination packet.
   For a unknown-unicast packet, if the DBRB has learnt the destination
   MAC address, it SHOULD convert the packet to unicast and set its M
   bit to 0. Otherwise, the multi-destination packet will continue to be
   flooded as multicast packet on the distribution tree. The DBRB
   chooses the new distribution tree by replacing the egress nickname
   with the new root RBridge nickname. The following sequence of events
   will occur:

  - RB2, when transitioning the packet from Level 1 to Level 2,
     replaces the ingress TRILL switch nickname with its own nickname,
     so replaces 27 with 2. RB2 also needs to replace the egress
     RBridge nickname with the L2 tree root RBridge nickname, say 2. In
     order to accommodate return traffic, RB2 records that S is
     attached to nickname 27 and SHOULD use ESADI protocol to
     synchronize this attachment information with other border RBridges
     (say RB20) in the area.

   - RB20, will receive the packet flooded on the L2 tree by RB2. It is
     important that RB20 does not transition this packet back to L1 as
     it does for a multicast packet normally received from another
     remote L1 area. RB20 should examine the ingress nickname of this
     packet. If this nickname is found to be a border RBridge nickname
     of the area {2,20}, RB2 must not forwarded the packet into this
     area.

   - The packet is flooded on the Level 2 tree to reach both RB3 and
     RB30. Suppose RB3 is the selected DBRB. The non-DBRB RB30 will
     drop the packet.

   - RB3, when forwarding into area {3,30}, replaces the egress
     nickname in the TRILL header with the root RBridge nickname, say
     3, of the distribution tree of L1 area {3,30}. (Here, the ingress
     nickname MAY be replaced with an area nickname selected from
     {2,20} as specified in Section 4.) Now suppose that RB27 has
     learned the location of D (attached to nickname 3), but RB3 does
     not know where D is. In that case, RB3 must turn the packet into a
     multi-destination packet and floods it on the distribution tree of
     L1 area {3,30}.

   - RB30, will receive the packet flooded on the L1 tree by RB3. It is
     important that RB30 does not transition this packet back to L2.
     RB30 should also examine the ingress nickname of this packet. If
     this nickname is found to be an L2 border RBridge nickname, RB30
     must not transition the packet back to L2.

   - The multicast listener RB44, when decapsulating the received
     packet, learns that S is attached to nickname 2, which is one of
     the area nicknames of the ingress.

4. Per-flow Load Balancing

   Area border RBridges perform ingress/egress nickname replacement when
   they transition TRILL data packets between Level 1 and Level 2. This
   nickname replacement enables the per-flow load balance which is
   specified as follows.

4.1. Ingress Nickname Replacement

   When a TRILL data packet from other areas arrives at an area border
   RBridge, this RBridge MAY select one area nickname of the ingress to
   replace the ingress nickname of the packet. The selection is simply
   based on a pseudorandom algorithm as defined in Section 5.3 of
   [RFC7357]. With the random ingress nickname replacement, the border
   RBridge actually achieves a per-flow load balance for returning
   traffic.

All area border RBridges in an L1 area MUST agree on the same
pseudorandom algorithm. The source MAC address, ingress area
nicknames, egress area nicknames and the Data Label of the received
TRILL data packet are candidate factors of the input of this
pseudorandom algorithm. Note that the value of the destination MAC
address SHOULD be excluded from the input of this pseudorandom
algorithm, otherwise the egress RBridge will see one source MAC
address flip flopping among multiple ingress RBridges.

4.2. Egress Nickname Replacement

When a TRILL data packet originated from the area arrives at an area
border RBridge, this RBridge MAY select one area nickname of the
egress to replace the egress nickname of the packet. By default, it
SHOULD choose the egress area border RBridge with the least cost
route to reach. The pseudorandom algorithm as defined in Section 5.3
of [RFC7357] may be used as well. In that case, however, the ingress
area border RBridge may take the non-least-cost Level 2 route to
forward the TRILL data packet to the egress area border RBridge.

5. Protocol Extensions for Discovery

5.1. Discovery of Border RBridges in L1

The following Level 1 Border RBridge APPsub-TLV will be included in
an E-L1FS FS-LSP fragment zero [RFC7180bis] as an APPsub-TLV of the
TRILL GENINFO-TLV. Through listening to this Appsub-TLV, an area
border RBridge discovers all other area border RBridges in this area.

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type = L1-BORDER-RBRIDGE      | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Length                        | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Sender Nickname               | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

o Type: Level 1 Border RBridge (TRILL APPsub-TLV type tbd1)

o Length: 2

o Sender Nickname: The nickname the originating IS will use as the
   L1 Border RBridge nickname. This field is useful because the
   originating IS might own multiple nicknames.

5.2. Discovery of Border RBridge Sets in L2

The following APPsub-TLV will be included in an E-L2FS FS-LSP

fragment zero [RFC7180bis] as an APPsub-TLV of the TRILL GENINFO-TLV.
Through listening to this APPsub-TLV in L2, an area border RBridge
discovers all groups of L1 border RBridges and each such group
identifies an area.

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type = L1-BORDER-RB-GROUP     | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Length                        | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| L1 Border RBridge Nickname 1  | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| ...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| L1 Border RBridge Nickname k  | (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

o Type: Level 1 Border RBridge Group (TRILL APPsub-TLV type tbd2)

o Length: 2 * k. If length is not a multiple of 2, the APPsub-TLV is
   corrupt and MUST be ignored.

o L1 Border RBridge Nickname: The nickname that an area border
   RBridge uses as the L1 Border RBridge nickname. The L1-BORDER-RB-
   GROUP TLV generated by an area border RBridge MUST include all L1
   Border RBridge nicknames of the area. It's RECOMMENDED that these
   k nicknames are ordered in ascending order according to the 2-
   octet nickname considered as an unsigned integer.

When an L1 area is partitioned [MultiL], border RBridges will re-
discover each other in both L1 and L2 through exchanging LSPs. In L2,
the set of border RBridge nicknames for this splitting area will
change. Border RBridges that detect such a change MUST flush the
reach-ability information associated to any RBridge nickname from
this changing set.

6. One Border RBridge Connects Multiple Areas

It's possible that one border RBridge (say RB1) connects multiple L1
areas. RB1 SHOULD use a single area nickname for all these areas.

Nicknames used within one of these areas can be reused within other
areas. It's important that packets destined to those duplicated
nicknames are sent to the right area. Since these areas are connected
to form a layer 2 network, duplicated {MAC, Data Label} across these
areas ought not occur. Now suppose a TRILL data packet arrives at the
area border nickname of RB1. For a unicast packet, RB1 can lookup the
{MAC, Data Label} entry in its MAC table to identify the right

destination area (i.e., the outgoing interface) and the egress
RBridge's nickname. For a multicast packet: suppose RB1 is not the
DBRB, RB1 will not transition the packet; otherwise, RB1 is the DBRB,

- if this packet is originated from an area out of the connected
  areas, RB1 should replicate this packet and flood it on the proper
  Level 1 trees of all the areas in which it acts as the DBRB.

- if the packet is originated from one of the connected areas, RB1
  should replicate the packet it receives from the Level 1 tree and
  flood it on other proper Level 1 trees of all the areas in which
  it acts as the DBRB except the originating area (i.e., the area
  connected to the incoming interface). RB1 may also receive the
  replication of the packet from the Level 2 tree. This replication
  must be dropped by RB1.

7. E-L1FS/E-L2FS Backwards Compatibility

All Level 2 RBridges MUST support E-L2FS [RFC7356] [rfc7180bis]. The
Extended TLVs defined in Section 5 are to be used in Extended Level
1/2 Flooding Scope (E-L1FS/E-L2FS) PDUs. Area border RBridges MUST
support both E-L1FS and E-L2FS. RBridges that do not support either
E-L1FS or E-L2FS cannot serve as area border RBridges but they can
well appear in an L1 area acting as non-area-border RBridges.

8. Security Considerations

For general TRILL Security Considerations, see [RFC6325].

The newly defined TRILL APPsub-TLVs in Section 5 are transported in
IS-IS PDUs whose authenticity can be enforced using regular IS-IS
security mechanism [IS-IS] [RFC5310]. This document raises no new
security issues for IS-IS.

9. IANA Considerations

9.1. TRILL APPsub-TLVs

IANA is requested to allocate two new types under the TRILL GENINFO
TLV [RFC7357] for the TRILL APPsub-TLVs defined in Section 5. The
following entries are added to the "TRILL APPsub-TLV Types under IS-
IS TLV 251 Application Identifier 1" Registry on the TRILL Parameters
IANA web page.

```
   Type        Name                    Reference
   ---------   ----                    ---------
   tbd1[256]   L1-BORDER-RBRIDGE       [This document]
   tbd2[257]   L1-BORDER-RB-GROUP      [This document]
```

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
          Requirement Levels", BCP 14, RFC 2119, DOI
          10.17487/RFC2119, March 1997, <http://www.rfc-
          editor.org/info/rfc2119>.

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

[RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding
          Scope Link State PDUs (LSPs)", RFC 7356, DOI
          10.17487/RFC7356, September 2014, <http://www.rfc-
          editor.org/info/rfc7356>.

[RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O.
          Stokes, "Transparent Interconnection of Lots of Links
          (TRILL): End Station Address Distribution Information
          (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357,
          September 2014, <http://www.rfc-editor.org/info/rfc7357>.

10.2. Informative References

[IS-IS]   International Organization for Standardization, ISO/IEC
          10589:2002, "Information technology -- Telecommunications
          and information exchange between systems -- Intermediate
          System to Intermediate System intra-domain routeing
          information exchange protocol for use in conjunction with
          the protocol for providing the connectionless-mode network
          service", ISO 8473, Second Edition, November 2002.

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R.,
          and M. Fanto, "IS-IS Generic Cryptographic Authentication",
          RFC 5310, DOI 10.17487/RFC5310, February 2009,
          <http://www.rfc-editor.org/info/rfc5310>.

[RFC7180bis] D. Eastlake, M. Zhang, et al, "TRILL: Clarifications,
          Corrections, and Updates", draft-ietf-trill-rfc7180bis,
          work in progress.

[MultiL] Perlman, R., Eastlake, D., et al, "Alternatives for
          Multilevel TRILL (Transparent Interconnection of Lots of
          Links)", draft-ietf-trill-rbridge-multilevel, work in
          progress.

Appendix A. Clarifications

A.1. Level Transition

   It's possible that an L1 RBridge is only reachable from a non-DBRB
   RBridge. If this non-DBRB RBridge refrains from Level transition, the
   question is, how can a multicast packet reach this L1 RBridge? The
   answer is, it will be reached after the DBRB performs the Level
   transition and floods the packet using an L1 distribution tree.

   Take the following figure as an example. RB77 is reachable from the
   border RBridge RB30 while RB3 is the DBRB. RB3 transitions the
   multicast packet into L1 and floods the packet on the distribution
   tree rooted from RB3. This packet will finally flooded to RB77 via
   RB30.

```
            Area{3,30}
            +--------------+           (root) RB3 o
            |              |                        \
       -RB3 |              |                          o RB30
        |  |              |                          /
       -RB30-RB77          |                   RB77 o
            +--------------+


            Example Topology              L1 Tree
```

   In the above example, the multicast packet is forwarded along a non-
   optimal path. A possible improvement is to have RB3 configured not to
   belong to this area. In this way, RB30 will surely act as the DBRB to
   do the Level transition.

Author's Addresses

   Mingui Zhang
   Huawei Technologies
   No. 156 Beiqing Rd. Haidian District
   Beijing 100095
   China

   Email: zhangmingui@huawei.com

   Donald E. Eastlake, 3rd
   Huawei Technologies
   155 Beaver Street
   Milford, MA 01757
   United States

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

   Radia Perlman
   EMC
   2010 256th Avenue NE, #200
   Bellevue, WA 98007
   United States

   Email: radia@alum.mit.edu

   Margaret Cullen
   Painless Security
   356 Abbott Street
   North Andover, MA 01845
   United States

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

   Hongjun Zhai
   Jinling Institute of Technology
   99 Hongjing Avenue, Jiangning District
   Nanjing, Jiangsu 211169
   China

   Email: honjun.zhai@tom.com