TRILL Working Group
INTERNET-DRAFT
Intended status: Informational

Radia Perlman
EMC
Donald Eastlake
Mingui Zhang
Huawei
Anoop Ghanwani
Dell
Hongjun Zhai
JIT

Expires: February 27, 2017

August 28, 2016

Alternatives for Multilevel TRILL (Transparent Interconnection of Lots of Links) <draft-ietf-trill-rbridge-multilevel-04.txt>

Abstract

Extending TRILL to multiple levels has challenges that are not addressed by the already-existing capability of IS-IS to have multiple levels. One issue is with the handling of multi-destination packet distribution trees. Another issue is with TRILL switch nicknames. There have been two proposed approaches. One approach, which we refer to 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 we refer to as the "aggregated nickname" approach, involves hiding the nicknames within areas, allowing nicknames to be reused in different areas, by having the border TRILL switches rewrite the nickname fields when entering or leaving an area. Each of those approaches has advantages and disadvantages.

This informational document suggests allowing a choice of approach in each area. This allows the simplicity of the unique nickname approach in installations in which there is no danger of running out of nicknames and allows the complexity of hiding the nicknames in an area to be phased into larger installations on a per-area basis.

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79. Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list <trill@ietf.org>.

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/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....4
1.1 TRILL Scalability Issues....4
1.2 Improvements Due to Multilevel....5
1.3 Unique and Aggregated Nicknames....6
1.3 More on Areas....7
1.4 Terminology and Acronyms....7

2. Multilevel TRILL Issues....9
2.1 Non-zero Area Addresses....10
2.2 Aggregated versus Unique Nicknames....10
2.2.1 More Details on Unique Nicknames....11
2.2.2 More Details on Aggregated Nicknames....12
2.2.2.1 Border Learning Aggregated Nicknames....13
2.2.2.2 Swap Nickname Field Aggregated Nicknames....15
2.2.2.3 Comparison....15
2.3 Building Multi-Area Trees....16
2.4 The RPF Check for Trees....16
2.5 Area Nickname Acquisition....17
2.6 Link State Representation of Areas....17

3. Area Partition....19

4. Multi-Destination Scope....20
4.1 Unicast to Multi-destination Conversions....20
4.1.1 New Tree Encoding....21
4.2 Selective Broadcast Domain Reduction....21

5. Co-Existence with Old TRILL switches....23
6. Multi-Access Links with End Stations....24
7. Summary....25

8. Security Considerations....26
9. IANA Considerations....26

Normative References....27
Informative References....27
Acknowledgements....29

Authors' Addresses....30

1. Introduction

The IETF TRILL (Transparent Interconnection of Lot of Links or Tunneled Routing in the Link Layer) protocol [RFC6325] [RFC7177] [RFC7780] provides optimal pair-wise data routing without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic in networks with arbitrary topology and link technology, including multi-access links. TRILL accomplishes this by using IS-IS (Intermediate System to Intermediate System [IS-IS] [RFC7176]) link state routing in conjunction with a header that includes a hop count. The design supports data labels (VLANs and Fine Grained Labels [RFC7172]) and optimization of the distribution of multi-destination data based on VLANs and multicast groups. Devices that implement TRILL are called TRILL Switches or RBridges.

Familiarity with [IS-IS], [RFC6325], and [RFC7780] is assumed in this document.

1.1 TRILL Scalability Issues

There are multiple issues that might limit the scalability of a TRILL-based network as follows:

1. The routing computation load
2. The volatility of the link state database (LSDB) creating too much control traffic
3. The volatility of the LSDB causing the TRILL network to be in an unconverged state too much of the time
4. The size of the LSDB
5. The limit of the number of TRILL switches, due to the 16-bit nickname space
6. The traffic due to upper layer protocols use of broadcast and multicast
7. The size of the end node learning table (the table that remembers (egress TRILL switch, label/MAC) pairs)

As discussed below, extending TRILL IS-IS to be multilevel (hierarchical) helps with all but the last of these issues.

IS-IS was designed to be multilevel [IS-IS]. A network can be partitioned into "areas". Routing within an area is known as "Level 1 routing". Routing between areas is known as "Level 2 routing". The Level 2 IS-IS network consists of Level 2 routers and links between the Level 2 routers. Level 2 routers may participate in one or more Level 1 areas, in addition to their role as Level 2 routers.

Each area is connected to Level 2 through one or more "border

routers", which participate both as a router inside the area, and as a router inside the Level 2 "area". Care must be taken that it is clear, when transitioning multi-destination packets between Level 2 and a Level 1 area in either direction, that exactly one border TRILL switch will transition a particular data packet between the levels or else duplication or loss of traffic can occur.

1.2 Improvements Due to Multilevel

Partitioning the network into areas solves the first four scalability issues described above as listed below. In this section, N indicates the number of TRILL switches in a TRILL campus.

1. The routing computation load

   The optimized computational effort to calculate least cost routes at a TRILL switch in a single level campus is on the order of N*log(N). In an optimized multi-level campus, it is on the order of sqrt(N)*log(N). So, for example, assuming N is 3,000, the level of computational effort would be reduced by about a factor of 50.

2. The volatility of the LSDB creating too much control traffic

   The rate of LSDB changes would be approximately proportional to the number of routers/links in the TRILL campus for a single level campus. With an optimized multi-level campus, each area would have about sqrt(N) routers reducing volatility by about a factor of sqrt(N).

3. The volatility of the LSDB causing the TRILL network to be in an unconverged state too much of the time

   With the simplifying assumption that routing converges after each change before the next change, the fraction of time that routing is unconverged is proportional to the product of the volatility and the convergence time and the convergence time is approximately proportional to the computation involved at each router. Thus, based on these simplifying assumptions, the fraction of time routing at a router is not converged with the network would improve, in going from single to multi-level, by about a factor of N.

4. The size of the LSDB

   The size of the LSDB is also approximately proportional to the number of routers/links and so, as with item 2 above, should improve by about a factor of sqrt(N) in going from single to

multi-level.

Problem #6 in Section 1.1, namely, the traffic due to upper layer protocols use of broadcast and multicast, can be addressed by introducing a locally-scoped multi-destination delivery, limited to an area or a single link. See further discussion in Section 4.2.

Problem #5 in Section 1.1, namely, the limit of the number of TRILL switches, due to the 16-bit nickname space, will only be addressed with the aggregated nickname approach. Since the aggregated nickname approach requires some complexity in the border TRILL switches (for rewriting the nicknames in the TRILL header), the design in this document allows a campus with a mixture of unique-nickname areas, and aggregated-nickname areas. Nicknames must be unique across all Level 2 and unique-nickname area TRILL switches, whereas nicknames inside an aggregated-nickname area are visible only inside the area. Nicknames inside an aggregated-nickname area must not conflict with nicknames visible in Level 2 (which includes all nicknames inside unique nickname areas), but the nicknames inside an aggregated-nickname area may be the same as nicknames used within other aggregated-nickname areas.

TRILL switches within an area need not be aware of whether they are in an aggregated nickname area or a unique nickname area. The border TRILL switches in area A1 will claim, in their LSP inside area A1, which nicknames (or nickname ranges) are not available for choosing as nicknames by area A1 TRILL switches.

1.3 Unique and Aggregated Nicknames

We describe two alternatives for hierarchical or multilevel TRILL. One we call the "unique nickname" alternative. The other we call the "aggregated nickname" alternative. In the aggregated nickname alternative, border TRILL switches replace either the ingress or egress nickname field in the TRILL header of unicast packets with an aggregated nickname representing an entire area.

The unique nickname alternative has the advantage that border TRILL switches are simpler and do not need to do TRILL Header nickname modification. It also simplifies testing and maintenance operations that originate in one area and terminate in a different area.

The aggregated nickname alternative has the following advantages:

- it solves problem #5 above, the 16-bit nickname limit, in a simple way,
- it lessens the amount of inter-area routing information that must be passed in IS-IS, and
- it logically reduces the RPF (Reverse Path Forwarding) Check

information (since only the area nickname needs to appear,

rather than all the ingress TRILL switches in that area).

In both cases, it is possible and advantageous to compute multi-destination data packet distribution trees such that the portion computed within a given area is rooted within that area.

1.3 More on Areas

Each area is configured with an "area address", which is advertised in IS-IS messages, so as to avoid accidentally interconnecting areas. Although the area address had other purposes in CLNP (Connectionless Network Layer Protocol, IS-IS was originally designed for CLNP/DECnet), for TRILL the only purpose of the area address would be to avoid accidentally interconnecting areas.

Currently, the TRILL specification says that the area address must be zero. If we change the specification so that the area address value of zero is just a default, then most of IS-IS multilevel machinery works as originally designed. However, there are TRILL-specific issues, which we address below in this document.

1.4 Terminology and Acronyms

This document generally uses the acronyms defined in [RFC6325] plus the additional acronym DBRB. However, for ease of reference, most acronyms used are listed here:

CLNP - ConnectionLess Network Protocol

DECnet - a proprietary routing protocol that was used by Digital Equipment Corporation. "DECnet Phase 5" was the origin of IS-IS.

Data Label - VLAN or Fine Grained Label [RFC7172]

DBRB - Designated Border RBridge

ESADI - End Station Address Distribution Information

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

LSDB - Link State Data Base

LSP - Link State PDU

PDU - Protocol Data Unit

RBridge - Routing Bridge, an alternative name for a TRILL switch

RPF - Reverse Path Forwarding

TLV - Type Length Value

TRILL - Transparent Interconnection of Lots of Links or Tunneled Routing in the Link Layer [RFC6325] [RFC7780]

TRILL switch - a device that implements the TRILL protocol [RFC6325] [RFC7780], sometimes called an RBridge

VLAN - Virtual Local Area Network

2. Multilevel TRILL Issues

The TRILL-specific issues introduced by multilevel include the following:

a. Configuration of non-zero area addresses, encoding them in IS-IS PDUs, and possibly interworking with old TRILL switches that do not understand non-zero area addresses.

   See Section 2.1.

b. Nickname management.

   See Sections 2.5 and 2.2.

c. Advertisement of pruning information (Data Label reachability, IP multicast addresses) across areas.

   Distribution tree pruning information is only an optimization, as long as multi-destination packets are not prematurely pruned. For instance, border TRILL switches could advertise they can reach all possible Data Labels, and have an IP multicast router attached. This would cause all multi-destination traffic to be transmitted to border TRILL switches, and possibly pruned there, when the traffic could have been pruned earlier based on Data Label or multicast group if border TRILL switches advertised more detailed Data Label and/or multicast listener and multicast router attachment information.

d. Computation of distribution trees across areas for multi-destination data.

   See Section 2.3.

e. Computation of RPF information for those distribution trees.

   See Section 2.4.

f. Computation of pruning information across areas.

   See Sections 2.3 and 2.6.

g. Compatibility, as much as practical, with existing, unmodified TRILL switches.

   The most important form of compatibility is with existing TRILL fast path hardware. Changes that require upgrade to the slow path firmware/software are more tolerable. Compatibility for the relatively small number of border TRILL switches is less

important than compatibility for non-border TRILL switches.

See Section 5.

2.1 Non-zero Area Addresses

The current TRILL base protocol specification [RFC6325] [RFC7177] [RFC7780] says that the area address in IS-IS must be zero. The purpose of the area address is to ensure that different areas are not accidentally merged. Furthermore, zero is an invalid area address for layer 3 IS-IS, so it was chosen as an additional safety mechanism to ensure that layer 3 IS-IS would not be confused with TRILL IS-IS. However, TRILL uses other techniques to avoid such confusion, such as different multicast addresses and Ethertypes on Ethernet [RFC6325], different PPP (Point-to-Point Protocol) code points on PPP [RFC6361], and the like. Thus, using an area address in TRILL that might be used in layer 3 IS-IS is not a problem.

Since current TRILL switches will reject any IS-IS messages with non-zero area addresses, the choices are as follows:

a.1 upgrade all TRILL switches that are to interoperate in a potentially multilevel environment to understand non-zero area addresses,
a.2 neighbors of old TRILL switches must remove the area address from IS-IS messages when talking to an old TRILL switch (which might break IS-IS security and/or cause inadvertent merging of areas),
a.3 ignore the problem of accidentally merging areas entirely, or
a.4 keep the fixed "area address" field as 0 in TRILL, and add a new, optional TLV for "area name" to Hellos that, if present, could be compared, by new TRILL switches, to prevent accidental area merging.

In principal, different solutions could be used in different areas but it would be much simpler to adopt one of these choices uniformly.

2.2 Aggregated versus Unique Nicknames

In the unique nickname alternative, all nicknames across the campus must be unique. In the aggregated nickname alternative, TRILL switch nicknames within an aggregated area are only of local significance, and the only nickname externally (outside that area) visible is the "area nickname" (or nicknames), which aggregates all the internal nicknames.

The unique nickname approach simplifies border TRILL switches.

The aggregated nickname approach eliminates the potential problem of

nickname exhaustion, minimizes the amount of nickname information that would need to be forwarded between areas, minimizes the size of the forwarding table, and simplifies RPF calculation and RPF information.

2.2.1 More Details on Unique Nicknames

With unique cross-area nicknames, it would be intractable to have a flat nickname space with TRILL switches in different areas contending for the same nicknames. Instead, each area would need to be configured with a block of nicknames. Either some TRILL switches would need to announce that all the nicknames other than that block are taken (to prevent the TRILL switches inside the area from choosing nicknames outside the area's nickname block), or a new TLV would be needed to announce the allowable nicknames, and all TRILL switches in the area would need to understand that new TLV.

Currently the encoding of nickname information in TLVs is by listing of individual nicknames; this would make it painful for a border TRILL switch to announce into an area that it is holding all other nicknames to limit the nicknames available within that area. The information could be encoded as ranges of nicknames to make this somewhat manageable by specifying a new TLV similar to the Nickname Flags APPsubTLV specified in [RFC7780] but providing flags for blocks of nicknames rather than single nicknames; however, such a new TLV would not be intelligible to old TRILL switches.

There is also an issue with the unique nicknames approach in building distribution trees, as follows:

> With unique nicknames in the TRILL campus and TRILL header nicknames not rewritten by the border TRILL switches, there would have to be globally known nicknames for the trees. Suppose there are k trees. For all of the trees with nicknames located outside an area, the local trees would be rooted at a border TRILL switch or switches. Therefore, there would be either no splitting of multi-destination traffic within the area or restricted splitting of multi-destination traffic between trees rooted at a highly restricted set of TRILL switches.
>
> As an alternative, just the "egress nickname" field of multi-destination TRILL Data packets could be mapped at the border, leaving known unicast packets un-mapped. However, this surrenders much of the unique nickname advantage of simpler border TRILL switches.

Scaling to a very large campus with unique nicknames might exhaust

the 16-bit TRILL nicknames space particularly if (1) additional

nicknames are consumed to support active-active end station groups at the TRILL edge using the techniques standardized in [RFC7781] and (2) use of the nickname space is less efficient due to the allocation of, for example, power-of-two size blocks of nicknames to areas in the same way that use of the IP address space is made less efficient by hierarchical allocation (see [RFC3194]). One method to avoid nickname exhaustion might be to expand nicknames to 24 bits; however, that technique would require TRILL message format changes and that all TRILL switches in the campus understand larger nicknames.

2.2.2 More Details on Aggregated Nicknames

The aggregated nickname approach enables passing far less nickname information. It works as follows, assuming both the source and destination areas are using aggregated nicknames:

There are two ways areas could be identified.

> One method would be to assign each area a 16-bit nickname. This would not be the nickname of any actual TRILL switch. Instead, it would be the nickname of the area itself. Border TRILL switches would know the area nickname for their own area(s). For an example of a more specific multilevel proposal using unique nicknames, see [DraftUnique].
>
> Alternatively, areas could be identified by the set of nicknames that identify the border routers for that area. (See [SingleName] for a multilevel proposal using such a set of nicknames.)

The TRILL Header nickname fields in TRILL Data packets being transported through a multilevel TRILL campus with aggregated nicknames are as follows:

- When both the ingress and egress TRILL switches are in the same area, there need be no change from the existing base TRILL protocol standard in the TRILL Header nickname fields.

- When being transported in Level 2, the ingress nickname is a nickname of the ingress TRILL switch's area while the egress nickname is either a nickname of the egress TRILL switch's area or a tree nickname.

- When being transported from Level 1 to Level 2, the ingress nickname is the nickname of the ingress TRILL switch itself while the egress nickname is either a nickname for the area of the egress TRILL switch or a tree nickname.

- When being transported from Level 2 to Level 1, the ingress

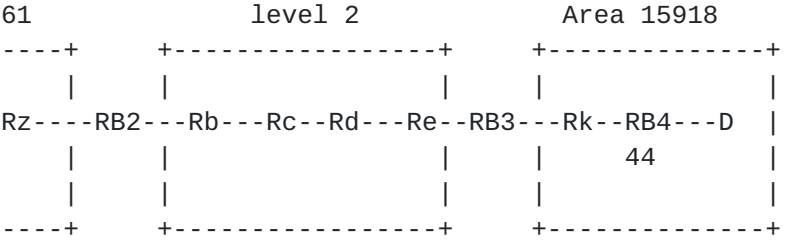
nickname is a nickname for the ingress TRILL switch's area while the egress nickname is either the nickname of the egress TRILL switch itself or a tree nickname.

There are two variations of the aggregated nickname approach. The first is the Border Learning approach, which is described in Section 2.2.2.1. The second is the Swap Nickname Field approach, which is described in Section 2.2.2.2. Section 2.2.2.3 compares the advantages and disadvantages of these two variations of the aggregated nickname approach.

2.2.2.1 Border Learning Aggregated Nicknames

This section provides an illustrative example and description of the border learning variation of aggregated nicknames where a single nickname is used to identify an area.

In the following picture, RB2 and RB3 are area border TRILL switches (RBridges). A source S is attached to RB1. The two areas have nicknames 15961 and 15918, respectively. RB1 has a nickname, say 27, and RB4 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 aggregated areas).



Let's say that S transmits a frame to destination D, which is connected to RB4, and let's say that D's location has already been learned by the relevant TRILL switches. These relevant switches have learned the following:

1) RB1 has learned that D is connected to nickname 15918
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.

- RB1 encapsulates with a TRILL header with ingress RBridge = 27, and egress = 15918 producing a TRILL Data packet.

- RB2 has announced in the Level 1 IS-IS instance in area 15961, that it is attached to all the area nicknames, including 15918. Therefore, IS-IS routes the packet to RB2. Alternatively, if a distinguished range of nicknames is used for Level 2, Level 1 TRILL switches seeing such an egress nickname will know to route to the nearest border router, which can be indicated by the IS-IS attached bit.

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with the area nickname, so replaces 27 with 15961. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 15961, and the egress RBridge field will be 15918. Also RB2 learns that S is attached to nickname 27 in area 15961 to accommodate return traffic.

- The packet is forwarded through Level 2, to RB3, which has advertised, in Level 2, reachability to the nickname 15918.

- RB3, when forwarding into area 15918, replaces the egress nickname in the TRILL header with RB4's nickname (44). So, within the destination area, the ingress nickname will be 15961 and the egress nickname will be 44.

- RB4, when decapsulating, learns that S is attached to nickname 15961, which is the area nickname of the ingress.

Now suppose that D's location has not been learned by RB1 and/or RB3. What will happen, as it would in TRILL today, is that RB1 will forward the packet as multi-destination, choosing a tree. As the multi-destination packet transitions into Level 2, RB2 replaces the ingress nickname with the area nickname. If RB1 does not know the location of D, the packet must be flooded, subject to possible pruning, in Level 2 and, subject to possible pruning, from Level 2 into every Level 1 area that it reaches on the Level 2 distribution tree.

Now suppose that RB1 has learned the location of D (attached to nickname 15918), but RB3 does not know where D is. In that case, RB3 must turn the packet into a multi-destination packet within area 15918. In this case, care must be taken so that in the case in which RB3 is not the Designated transitioner between Level 2 and its area for that multi-destination packet, but was on the unicast path, that border TRILL switch in that area does not forward the now multi-destination packet back into Level 2. Therefore, it would be desirable to have a marking, somehow, that indicates the scope of this packet's distribution to be "only this area" (see also Section 4).

In cases where there are multiple transitioners for unicast packets, the border learning mode of operation requires that the address

learning between them be shared by some protocol such as running ESADI [RFC7357] for all Data Labels of interest to avoid excessive unknown unicast flooding.

The potential issue described at the end of Section 2.2.1 with trees in the unique nickname alternative is eliminated with aggregated nicknames. With aggregated nicknames, each border TRILL switch that will transition multi-destination packets can have a mapping between Level 2 tree nicknames and Level 1 tree nicknames. There need not even be agreement about the total number of trees; just that the border TRILL switch have some mapping, and replace the egress TRILL switch nickname (the tree name) when transitioning levels.

2.2.2.2 Swap Nickname Field Aggregated Nicknames

As a variant, two additional fields could exist in TRILL Data packets we call the "ingress swap nickname field" and the "egress swap nickname field". The changes in the example above would be as follows:

- RB1 will have learned the area nickname of D and the TRILL switch nickname of RB4 to which D is attached. In encapsulating a frame to D, it puts an area nickname of D (15918) in the egress nickname field of the TRILL Header and puts a nickname of RB3 (44) in a egress swap nickname field.

- RB2 moves the ingress nickname to the ingress swap nickname field and inserts 15961, an area nickname for S, into the ingress nickname field.

- RB3 swaps the egress nickname and the egress swap nickname fields, which sets the egress nickname to 44.

- RB4 learns the correspondence between the source MAC/VLAN of S and the { ingress nickname, ingress swap nickname field } pair as it decapsulates and egresses the frame.

See [DraftAggregated] for a multilevel proposal using aggregated swap nicknames with a single nickname representing an area.

2.2.2.3 Comparison

The Border Learning variant described in Section 2.2.2.1 above minimizes the change in non-border TRILL switches but imposes the burden on border TRILL switches of learning and doing lookups in all

the end station MAC addresses within their area(s) that are used for

communication outside the area. This burden could be reduced by decreasing the area size and increasing the number of areas.

The Swap Nickname Field variant described in Section 2.2.2.2 eliminates the extra address learning burden on border TRILL switches but requires more extensive changes to non-border TRILL switches. In particular they must learn to associate both a TRILL switch nickname and an area nickname with end station MAC/label pairs (except for addresses that are local to their area).

The Swap Nickname Field alternative is more scalable but less backward compatible for non-border TRILL switches. It would be possible for border and other level 2 TRILL switches to support both Border Learning, for support of legacy Level 1 TRILL switches, and Swap Nickname, to support Level 1 TRILL switches that understood the Swap Nickname method.

2.3 Building Multi-Area Trees

It is easy to build a multi-area tree by building a tree in each area separately, (including the Level 2 "area"), and then having only a single border TRILL switch, say RBx, in each area, attach to the Level 2 area. RBx would forward all multi-destination packets between that area and Level 2.

People might find this unacceptable, however, because of the desire to path split (not always sending all multi-destination traffic through the same border TRILL switch).

This is the same issue as with multiple ingress TRILL switches injecting traffic from a pseudonode, and can be solved with the mechanism that was adopted for that purpose: the affinity TLV [RFC7783]. For each tree in the area, at most one border RB announces itself in an affinity TLV with that tree name.

2.4 The RPF Check for Trees

For multi-destination data originating locally in RBx's area, computation of the RPF check is done as today. For multi-destination packets originating outside RBx's area, computation of the RPF check must be done based on which one of the border TRILL switches (say RB1, RB2, or RB3) injected the packet into the area.

A TRILL switch, say RB4, located inside an area, must be able to know which of RB1, RB2, or RB3 transitioned the packet into the area from

Level 2 (or into Level 2 from an area).

This could be done based on having the DBRB announce the transitioner assignments to all the TRILL switches in the area, or the Affinity TLV mechanism given in [RFC7783], or the New Tree Encoding mechanism discussed in Section 4.1.1.

2.5 Area Nickname Acquisition

In the aggregated nickname alternative, each area must acquire a unique area nickname or can be identified by the set of border TRILL switches. It is probably simpler to allocate a block of nicknames (say, the top 4000) to either (1) represent areas and not specific TRILL switches or (2) used by border TRILL switches if the set of such border TRILL switches represent the area.

The nicknames used for area identification need to be advertised and acquired through Level 2.

Within an area, all the border TRILL switches can discover each other through the Level 1 link state database, by using the IS-IS attach bit or by explicitly advertising in their LSP "I am a border RBridge".

Of the border TRILL switches, one will have highest priority (say RB7). RB7 can dynamically participate, in Level 2, to acquire a nickname for identifying the area. Alternatively, RB7 could give the area a pseudonode IS-IS ID, such as RB7.5, within Level 2. So an area would appear, in Level 2, as a pseudonode and the pseudonode could participate, in Level 2, to acquire a nickname for the area.

Within Level 2, all the border TRILL switches for an area can advertise reachability to the area, which would mean connectivity to a nickname identifying the area.

2.6 Link State Representation of Areas

Within an area, say area A1, there is an election for the DBRB, (Designated Border RBridge), say RB1. This can be done through LSPs within area A1. The border TRILL switches announce themselves, together with their DBRB priority. (Note that the election of the DBRB cannot be done based on Hello messages, because the border TRILL switches are not necessarily physical neighbors of each other. They can, however, reach each other through connectivity within the area, which is why it will work to find each other through Level 1 LSPs.)

RB1 can acquire an area nickname (in the aggregated nickname

approach) and may give the area a pseudonode IS-IS ID (just like the

DRB would give a pseudonode IS-IS ID to a link) depending on how the area nickname is handled. RB1 advertises, in area A1, an area nickname that RB1 has acquired (and what the pseudonode IS-IS ID for the area is if needed).

Level 1 LSPs (possibly pseudonode) initiated by RB1 for the area include any information external to area A1 that should be input into area A1 (such as nicknames of external areas, or perhaps (in the unique nickname variant) all the nicknames of external TRILL switches in the TRILL campus and pruning information such as multicast listeners and labels). All the other border TRILL switches for the area announce (in their LSP) attachment to that area.

Within Level 2, RB1 generates a Level 2 LSP on behalf of the area. The same pseudonode ID could be used within Level 1 and Level 2, for the area. (There does not seem any reason why it would be useful for it to be different, but there's also no reason why it would need to be the same). Likewise, all the area A1 border TRILL switches would announce, in their Level 2 LSPs, connection to the area.

3. Area Partition

It is possible for an area to become partitioned, so that there is still a path from one section of the area to the other, but that path is via the Level 2 area.

With multilevel TRILL, an area will naturally break into two areas in this case.

Area addresses might be configured to ensure two areas are not inadvertently connected. Area addresses appear in Hellos and LSPs within the area. If two chunks, connected only via Level 2, were configured with the same area address, this would not cause any problems. (They would just operate as separate Level 1 areas.)

A more serious problem occurs if the Level 2 area is partitioned in such a way that it could be healed by using a path through a Level 1 area. TRILL will not attempt to solve this problem. Within the Level 1 area, a single border RBridge will be the DBRB, and will be in charge of deciding which (single) RBridge will transition any particular multi-destination packets between that area and Level 2. If the Level 2 area is partitioned, this will result in multi-destination data only reaching the portion of the TRILL campus reachable through the partition attached to the TRILL switch that transitions that packet. It will not cause a loop.

4. Multi-Destination Scope

There are at least two reasons it would be desirable to be able to mark a multi-destination packet with a scope that indicates the packet should not exit the area, as follows:

1. To address an issue in the border learning variant of the aggregated nickname alternative, when a unicast packet turns into a multi-destination packet when transitioning from Level 2 to Level 1, as discussed in Section 4.1.

2. To constrain the broadcast domain for certain discovery, directory, or service protocols as discussed in Section 4.2.

Multi-destination packet distribution scope restriction could be done in a number of ways. For example, there could be a flag in the packet that means "for this area only". However, the technique that might require the least change to TRILL switch fast path logic would be to indicate this in the egress nickname that designates the distribution tree being used. There could be two general tree nicknames for each tree, one being for distribution restricted to the area and the other being for multi-area trees. Or there would be a set of N (perhaps 16) special currently reserved nicknames used to specify the N highest priority trees but with the variation that if the special nickname is used for the tree, the packet is not transitioned between areas. Or one or more special trees could be built that were restricted to the local area.

4.1 Unicast to Multi-destination Conversions

In the border learning variant of the aggregated nickname alternative, the following situation may occur:

- a unicast packet might be known at the Level 1 to Level 2 transition and be forwarded as a unicast packet to the least cost border TRILL switch advertising connectivity to the destination area, but
- upon arriving at the border TRILL switch, it turns out to have an unknown destination { MAC, Data Label } pair.

In this case, the packet must be converted into a multi-destination packet and flooded in the destination area. However, if the border TRILL switch doing the conversion is not the border TRILL switch designated to transition the resulting multi-destination packet, there is the danger that the designated transitioner may pick up the packet and flood it back into Level 2 from which it may be flooded into multiple areas. This danger can be avoided by restricting any multi-destination packet that results from such a conversion to the

destination area through a flag in the packet or though distributing

it on a tree that is restricted to the area, or other techniques (see Section 4).

Alternatively, a multi-destination packet intended only for the area could be tunneled (within the area) to the RBridge RBx, that is the appointed transitioner for that form of packet (say, based on VLAN or FGL), with instructions that RBx only transmit the packet within the area, and RBx could initiate the multi-destination packet within the area. Since RBx introduced the packet, and is the only one allowed to transition that packet to Level 2, this would accomplish scoping of the packet to within the area. Since this case only occurs in the unusual case when unicast packets need to be turned into multi-destination as described above, the suboptimality of tunneling between the border TRILL switch that receives the unicast packet and the appointed level transitioner for that packet, would not be an issue.

4.1.1 New Tree Encoding

The current encoding, in a TRILL header, of a tree, is of the nickname of the tree root. This requires all 16 bits of the egress nickname field. TRILL could instead, for example, use the bottom 6 bits to encode the tree number (allowing 64 trees), leaving 10 bits to encode information such as:

- scope: a flag indicating whether it should be single area only, or entire campus
- border injector: an indicator of which of the k border TRILL switches injected this packet

If TRILL were to adopt this new encoding, any of the TRILL switches in an edge group could inject a multi-destination packet. This would require all TRILL switches to be changed to understand the new encoding for a tree, and it would require a TLV in the LSP to indicate which number each of the TRILL switches in an edge group would be.

4.2 Selective Broadcast Domain Reduction

There are a number of service, discovery, and directory protocols that, for convenience, are accessed via multicast or broadcast frames. Examples are DHCP, (Dynamic Host Configuration Protocol) the NetBIOS Service Location Protocol, and multicast DNS (Domain Name Service).

Some such protocols provide means to restrict distribution to an IP

subnet or equivalent to reduce size of the broadcast domain they are using and then provide a proxy that can be placed in that subnet to use unicast to access a service elsewhere. In cases where a proxy mechanism is not currently defined, it may be possible to create one that references a central server or cache. With multilevel TRILL, it is possible to construct very large IP subnets that could become saturated with multi-destination traffic of this type unless packets can be further restricted in their distribution. Such restricted distribution can be accomplished for some protocols, say protocol P, in a variety of ways including the following:

- Either (1) at all ingress TRILL switches in an area place all protocol P multi-destination packets on a distribution tree in such a way that the packets are restricted to the area or (2) at all border TRILL switches between that area and Level 2, detect protocol P multi-destination packets and do not transition them.

- Then place one, or a few for redundancy, protocol P proxies inside each area where protocol P may be in use. These proxies unicast protocol P requests or other messages to the actual campus server(s) for P. They also receive unicast responses or other messages from those servers and deliver them within the area via unicast, multicast, or broadcast as appropriate. (Such proxies would not be needed if it was acceptable for all protocol P traffic to be restricted to an area.)

While it might seem logical to connect the campus servers to TRILL switches in Level 2, they could be placed within one or more areas so that, in some cases, those areas might not require a local proxy server.

5. Co-Existence with Old TRILL switches

TRILL switches that are not multilevel aware may have a problem with calculating RPF Check and filtering information, since they would not be aware of the assignment of border TRILL switch transitioning.

A possible solution, as long as any old TRILL switches exist within an area, is to have the border TRILL switches elect a single DBRB (Designated Border RBridge), and have all inter-area traffic go through the DBRB (unicast as well as multi-destination). If that DBRB goes down, a new one will be elected, but at any one time, all inter-area traffic (unicast as well as multi-destination) would go through that one DRBR. However this eliminates load splitting at level transition.

6. Multi-Access Links with End Stations

Care must be taken, in the case where there are multiple TRILL switches on a link with end stations, that only one TRILL switch ingress/egress any given data packet from/to the end nodes. With existing, single level TRILL, this is done by electing a single Designated RBridge per link, which appoints a single Appointed Forwarder per VLAN [RFC7177] [RFC6439]. But suppose there are two (or more) TRILL switches on a link in different areas, say RB1 in area A1 and RB2 in area A2, and that the link contains end nodes. If RB1 and RB2 ignore each other's Hellos then they will both ingress/egress end node traffic from the link.

A simple rule is to use the TRILL switch or switches having the lowest numbered area, comparing area numbers as unsigned integers, to handle native traffic. This would automatically give multilevel-ignorant legacy TRILL switches, that would be using area number zero, highest priority for handling end stations, which they would try to do anyway.

Other methods are possible. For example doing the selection of Appointed Forwarders and of the TRILL switch in charge of that selection across all TRILL switches on the link regardless of area. However, a special case would then have to be made for legacy TRILL switches using area number zero.

Any of these techniques require multilevel aware RBridges to take actions based on Hellos from RBridges in other areas even though they will not form an adjacency with such RBridges.

7. Summary

This draft discusses issues and possible approaches to multilevel TRILL. The alternative using aggregated areas has significant advantages in terms of scalability over using campus wide unique nicknames, not just in avoiding nickname exhaustion, but by allowing RPF Checks to be aggregated based on an entire area. However, the alternative of using unique nicknames is simpler and avoids the changes in border TRILL switches required to support aggregated nicknames. It is possible to support both. For example, a TRILL campus could use simpler unique nicknames until scaling begins to cause problems and then start to introduce areas with aggregated nicknames.

Some issues are not difficult, such as dealing with partitioned areas. Other issues are more difficult, especially dealing with old TRILL switches.

8. Security Considerations

This informational document explores alternatives for the use of multilevel IS-IS in TRILL and generally does not consider security issues.

If aggregated nicknames are used in two areas that have the same area address and those areas merge, there is a possibility of a transient nickname collision that would not occur with unique nicknames. Such a collision could cause a data packet to be delivered to the wrong egress TRILL switch but it would still not be delivered to any end station in the wrong Data Label; thus such delivery would still conform to security policies.

For general TRILL Security Considerations, see [RFC6325].

9. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

Normative References

[IS-IS] - 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.

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

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

[RFC7177] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014, <http://www.rfc-editor.org/info/rfc7177>.

[RFC7780] - Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <http://www.rfc-editor.org/info/rfc7780>.

Informative References

[RFC3194] - Durand, A. and C. Huitema, "The H-Density Ratio for Address Assignment Efficiency An Update on the H ratio", RFC 3194, DOI 10.17487/RFC3194, November 2001, <http://www.rfc-editor.org/info/rfc3194>.

[RFC6361] - Carlson, J. and D. Eastlake 3rd, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC 6361, August 2011.

[RFC7172] - Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014

[RFC7176] - Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014.

[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, September 2014, <http://www.rfc-editor.org/info/rfc7357>.

[RFC7781] - Zhai, H., Senevirathne, T., Perlman, R., Zhang, M., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Pseudo-Nickname for Active-Active Access", RFC 7781, DOI 10.17487/RFC7781, February 2016, <http://www.rfc-editor.org/info/rfc7781>.

[RFC7783] - Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for Transparent Interconnection of Lots of Links (TRILL)", RFC 7783, DOI 10.17487/RFC7783, February 2016, <http://www.rfc-editor.org/info/rfc7783>.

[DraftAggregated] - Bhargav Bhikkaji, Balaji Venkat Venkataswami, Narayana Perumal Swamy, "Connecting Disparate Data Center/PBB/Campus TRILL sites using BGP", draft-balaji-trill-over-ip-multi-level, Work In Progress.

[DraftUnique] - M. Zhang, D. Eastlake, R. Perlman, M. Cullen, H. Zhai, D. Liu, "TRILL Multilevel Using Unique Nicknames", draft-ietf-trill-multilevel-unique-nickname, Work In Progress.

[SingleName] - Mingui Zhang, et. al, "Single Area Border RBridge Nickname for TRILL Multilevel", draft-ietf-trill-multilevel-single-nickname, Work in Progress.

Acknowledgements

The helpful comments and contributions of the following are hereby acknowledged:

David Michael Bond, Dino Farinacci, Sue Hares, Gayle Noble, Alexander Vainshtein, and Stig Venaas.

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

Authors' Addresses

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

EMail: radia@alum.mit.edu

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

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

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

EMail: zhangmingui@huawei.com

Anoop Ghanwani
Dell
5450 Great America Parkway
Santa Clara, CA 95054 USA

EMail: anoop@alumni.duke.edu

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

EMail: honjun.zhai@tom.com

Copyright and IPR Provisions

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

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