Internet Engineering Task Force (IETF)
Request for Comments: 8243
Category: Informational
ISSN: 2070-1721

R. Perlman
EMC
D. Eastlake 3rd
M. Zhang
Huawei
A. Ghanwani
Dell
H. Zhai
JIT

September 2017

Alternatives for Multilevel Transparent Interconnection of Lots of Links (TRILL)

Abstract

Although TRILL is based on IS-IS, which supports multilevel unicast routing, extending TRILL to multiple levels has challenges that are not addressed by the already-existing capabilities of IS-IS. One issue is with the handling of multi-destination packet distribution trees. Other issues are with TRILL switch nicknames. How are such nicknames allocated across a multilevel TRILL network? Do nicknames need to be unique across an entire multilevel TRILL network? Or can they merely be unique within each multilevel area?

This informational document enumerates and examines alternatives based on a number of factors including backward compatibility, simplicity, and scalability; it makes recommendations in some cases.

Status of This Memo

Table of Contents

1. Introduction

The IETF Transparent Interconnection of Lot of Links (TRILL) protocol [RFC6325] [RFC7177] [RFC7780] provides optimal pairwise 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 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 (FGLs) [RFC7172]) and optimization of the distribution of multi-destination data based on Data Label and multicast group. Devices that implement TRILL are called TRILL Switches or RBridges.

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

1.1. The Motivation for Multilevel

The primary motivation for multilevel TRILL is to improve scalability. The following issues might limit the scalability of a TRILL-based network:

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 (for further information on why this might be a problem, see Section 1.2.5)

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 / Media Access Control (MAC)) pairs)

As discussed below, extending TRILL IS-IS to be multilevel (hierarchical) can help with all of these issues except issue 7.

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 a 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; otherwise, duplication or loss of traffic can occur.

1.2. Improvements Due to Multilevel

Partitioning the network into areas directly solves the first four scalability issues listed above, as described in Sections 1.2.1 through 1.2.4. Multilevel also contributes to solving issues 5 and 6, as discussed in Sections 1.2.5 and 1.2.6, respectively.

In the subsections below, N indicates the number of TRILL switches in a TRILL campus. For simplicity, it is assumed that each TRILL switch has k links to other TRILL switches. An "optimized" multilevel campus is assumed to have Level 1 areas containing sqrt(N) switches.

1.2.1. The Routing Computation Load

The Dijkstra algorithm uses computational effort on the order of the number of links in a network (N*k) times the log of the number of nodes to calculate least cost routes at a router (Section 12.3.3 of [InterCon]). Thus, in a single-level TRILL campus, it is on the order of N*k*log(N). In an optimized multilevel campus, it is on the order of sqrt(N)*k*log(N). So, for example, assuming N is 3,000, the level of computational effort would be reduced by about a factor of 50.

1.2.2. LSDB Volatility Creating Too Much Control Traffic

The rate of LSDB changes is assumed to be approximately proportional to the number of routers and links in the TRILL campus or N*(1+k) for a single-level campus. With an optimized multilevel campus, each area would have about sqrt(N) routers and proportionately fewer links reducing the rate of LSDB changes by about a factor of sqrt(N).

1.2.3. LSDB Volatility Causing Too Much Time Unconverged

With the simplifying assumption that routing converges after each topology change before the next such change, the fraction of time that routing is unconverged is proportional to the product of the rate of change occurrence and the convergence time. The rate of topology changes per some arbitrary unit of time will be roughly proportional to the number of router and links (Section 1.2.2). The convergence time is approximately proportional to the computation involved at each router (Section 1.2.1). Thus, based on these simplifying assumptions, the time spent unconverged in a single-level network is proportional to (N*(1+k))*(N*k*log(N)) while that time for an optimized multilevel network would be proportional to (sqrt(N)*(1+k))*(sqrt(N)*k*log(N)). Thus, in changing to multilevel, the time spent unconverged, using these simplifying assumptions, is improved by about a factor of N.

1.2.4. The Size of the LSDB

The size of the LSDB, which consists primarily of information about routers (TRILL switches) and links, is also approximately proportional to the number of routers and links. So, as with item 2 in Section 1.2.2, it should improve by about a factor of sqrt(N) in going from single level to multilevel.

1.2.5. Nickname Limit

For many TRILL protocol purposes, RBridges are designated by 16-bit nicknames. While some values are reserved, this appears to provide enough nicknames to designated over 65,000 RBridges. However, this number is effectively reduced by the following two factors:

- Nicknames are consumed when pseudo-nicknames are used for the active-active connection of end stations. Using the techniques in [RFC7781], for example, could double the nickname consumption if there are extensive active-active edge groups connected to different sets of edge TRILL switch ports.

- There might be problems in multilevel campus-wide contention for single-nickname allocation of nicknames were allocated individually from a single pool for the entire campus. Thus, it seems likely that a hierarchical method would be chosen where blocks of nicknames are allocated at Level 2 to Level 1 areas and contention for a nickname by an RBridge in such a Level 1 area would be only within that area. Such hierarchical allocation leads to further effective loss of nicknames similar to the situation with IP addresses discussed in [RFC3194].

Even without the above effective reductions in nickname space, a very large multilevel TRILL campus, say one with 200 areas each containing 500 TRILL switches, could require 100,000 or more nicknames if all nicknames in the campus must be unique, which is clearly impossible with 16-bit nicknames.

This scaling limit, namely, 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 suggested design in this document allows a campus with a mixture of unique-nickname areas, and aggregated-nickname areas. Thus, a TRILL network could start using multilevel with the simpler unique nickname method and later add aggregated-nickname areas as a later stage of network growth.

With this design, nicknames must be unique across all Level 2 and unique-nickname area TRILL switches taken together; whereas nicknames inside an aggregated-nickname area are visible only inside that area. Nicknames inside an aggregated-nickname area must still 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 one or more other aggregated-nickname areas.

With the design suggested in this document, 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 indicate, in their LSP inside area A1, which nicknames (or nickname ranges) are or are not available to be chosen as nicknames by area A1 TRILL switches.

1.2.6. Multi-Destination Traffic

In many cases, scaling limits due to protocol use of broadcast and multicast can be addressed in a multilevel campus by introducing locally scoped multi-destination delivery, limited to an area or a single link. See further discussion of this issue in Section 4.2.

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 scaling issue 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.

For further discussion of the unique and aggregated-nickname alternatives, see Section 2.2.

1.4. 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. For TRILL, the only purpose of the area address would be to avoid accidentally interconnecting areas although the area address had other purposes in CLNP (ConnectionLess Network Protocol), IS-IS was originally designed for CLNP/DECnet.

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 IS-IS multilevel machinery works as originally designed. However, there are TRILL-specific issues, which we address in Section 2.1.

1.5. Terminology and Abbreviations

This document generally uses the abbreviations defined in [RFC6325] plus the additional abbreviation DBRB. However, for ease of reference, most abbreviations 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 DataBase

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 packets would not be confused with TRILL IS-IS packets. However, TRILL uses other techniques to avoid confusion on a link, 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. A simple solution would be a.1, with each TRILL switch using a dominant area nickname as its area address. For the unique-nickname alternative, the dominant nickname could be the lowest value nickname held by any border RBridge of the area. For the aggregated-nickname alternative, it could be the lowest nickname held by a border RBridge of the area or a nickname representing the area.

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-nickname 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 or allocate one or more blocks of nicknames. Either some TRILL switches would need to announce that all the nicknames other than those in blocks available to the area 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 or the prohibited 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. Painful means tens of thousands of individual nickname entries in the Level 1 LSDB. The information could be encoded as ranges of nicknames to make this manageable by specifying a new TLV similar to the Nickname Flags APPsub-TLV specified in [RFC7780] but providing flags for blocks of nicknames rather than single nicknames. Although this would require updating software, such a new TLV is the preferred method.

There is also an issue with the unique-nickname 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 unmapped. 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 and fast-path processing changes and all TRILL switches in the campus to 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 at least 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 [UNIQUE].

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 between different Level 1 areas in Level 2, the ingress nickname is a nickname of the ingress TRILL switch's area, whereas 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, whereas 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, whereas 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-nickname areas).

```
        Area 15961             level 2           Area 15918
    +-------------------+     +-----------------+     +--------------+
    |                   |     |                 |     |              |
    |  S--RB1---Rx--Rz----RB2---Rb---Rc--Rd---Re--RB3---Rk--RB4---D  |
    |     27            |     |                 |     |     44       |
    |                   |     |                 |     |              |
    +-------------------+     +-----------------+     +--------------+
```

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" [IS-IS].

- RB2, when transitioning the packet from Level 1 to Level 2, replaces the ingress TRILL switch nickname with the area nickname, so it 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 agreement 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

There is a variant possibility where two additional fields could exist in TRILL Data packets that could be called the "ingress swap nickname field" and the "egress swap nickname field". This variant is described below for completeness, but it would require fast-path hardware changes from the existing TRILL protocol. 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 an 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 [TRILL-IP] 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 minimizes the change in non-border TRILL switches, but it 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 it requires changes to the TRILL Data packet header and more extensive changes to non-border TRILL switches. In particular, with this alternative, non-border TRILL switches 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 Field, to support Level 1 TRILL switches that understood the Swap Nickname Field method based on variations in the TRILL header; however, this would be even more complex.

The requirement to change the TRILL header and fast-path processing to support the Swap Nickname Field variant make it impractical for the foreseeable future.

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.

However, people might find this unacceptable 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 it 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 using any one of a variety of mechanisms such as having the DBRB announce the transitioner assignments to all the TRILL switches in the area or using the Affinity sub-TLV mechanism given in [RFC7783] or with a 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 identifier, for example, by acquiring a unique area nickname or by using an identifier based on the area’s 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) be 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 LSDB, by using the IS-IS "attached bit" [IS-IS] 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, 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 Designated RBridge (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 as described above.

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 might 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 an 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.

While there are a number of advantages to this technique, it requires fast-path logic changes; thus, its deployment is not practical at this time. It is included here for completeness.

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, the NetBIOS Service Location Protocol, and multicast DNS.

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 they 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. Coexistence 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 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 one or more end stations, keeping in mind that end stations are TRILL ignorant. In particular, it is essential that only one TRILL switch ingress/egress any given data packet from/ to an end station so that connectivity is provided to that end station without duplicating end-station data and that loops are not formed due to one TRILL switch egressing data in native form (i.e., with no TRILL header) and having that data re-ingressed by another TRILL switch on the link.

With existing, single-level TRILL, this is done by electing a single DRB per link, which appoints a single Appointed Forwarder per VLAN [RFC7177] [RFC8139]. This mechanism depends on the RBridges establishing adjacency. 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, as shown below; and suppose that the link also has one or more end stations attached. If RB1 and RB2 ignore each other’s Hellos because they are in different areas, as they are required to do under normal IS-IS PDU processing rules, then they will not form an adjacency. If they are not adjacent, they will ignore each other for the Appointed Forwarder mechanism and will both ingress/egress end-station traffic on the link causing loops and duplication.

The problem is not avoiding adjacency or avoiding TRILL-Data-packet transfer between RB1 and RB2; the area address mechanism of IS-IS or possibly the use of topology constraints (or the like) does that quite well. The problem stems from end stations being TRILL ignorant; therefore, care must be taken so that multiple RBridges on a link do not ingress the same frame originated by an end station and so that an RBridge does not ingress a native frame egressed by a different RBridge because the RBridge mistakes the frame for a frame originated by an end station.

```
      +--------------------------------------------+
      |                   Level 2                  |
      +----------+---------------------+-----------+
      |  Area A1 |                     |  Area A2  |
      |   +---+  |                     |   +---+   |
      |   |RB1|  |                     |   |RB2|   |
      |   +-+-+  |                     |   +-+-+   |
      |     |    |                     |     |     |
      +-----|----+                     +-----|-----+
            |                                |
          --+---------+-------------+--------+-- Link
                      |             |
               +------+------+   +--+----------+
               | End Station |   | End Station |
               +-------------+   +-------------+
```

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

Other methods are possible. For example, making the selection of the Appointed Forwarders and 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.

These techniques require multilevel-aware TRILL switches to take actions based on Hellos from RBridges in other areas, even though they will not form an adjacency with such RBridges. However, the action is quite simple in the preferred case: if a TRILL switch sees Hellos from lower-numbered areas, then they would not act as an Appointed Forwarder on the link until the Hello timer for such Hellos had expired.

7. Summary

This document describes potential scaling issues in TRILL and possible approaches to multilevel TRILL as a solution or element of a solution to most of them.

The alternative using aggregated-nickname areas in multilevel TRILL 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 multilevel TRILL issues are not difficult, such as dealing with partitioned areas. Other issues are more difficult, especially dealing with old TRILL switches that are multilevel ignorant.

8. Security Considerations

This informational document explores alternatives for the design 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 does not require any IANA actions.

10. References

10.1. Normative References

[IS-IS] International Organization for Standardization, "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)", ISO/IEC 10589:2002, Second Edition, November 2002.

[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, <https://www.rfc-editor.org/info/rfc6325>.

[RFC7177] Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, DOI 10.17487/RFC7177, May 2014, <https://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, <https://www.rfc-editor.org/info/rfc7780>.

[RFC8139] Eastlake 3rd, D., Li, Y., Umair, M., Banerjee, A., and F. Hu, "Transparent Interconnection of Lots of Links (TRILL): Appointed Forwarders", RFC 8139, DOI 10.17487/RFC8139, June 2017, <https://www.rfc-editor.org/info/rfc8139>.

10.2. 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, <https://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, DOI 10.17487/RFC6361, August 2011, <https://www.rfc-editor.org/info/rfc6361>.

[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, DOI 10.17487/RFC7172, May 2014, <https://www.rfc-editor.org/info/rfc7172>.

[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, DOI 10.17487/RFC7176, May 2014, <https://www.rfc-editor.org/info/rfc7176>.

[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, <https://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, <https://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, <https://www.rfc-editor.org/info/rfc7783>.

[InterCon] Perlman, R., "Interconnection: Bridges, Routers, Switches, and Internetworking Protocols", Addison Wesley Longman, Second Edition, Chapter 3, 1999.

[TRILL-IP] Bhikkaji, B., Venkataswami, B., Mahadevan, R., Sundaram, S., and N. Swamy, "Connecting Disparate Data Center/PBB/ Campus TRILL sites using BGP", Work in Progress, draft-balaji-trill-over-ip-multi-level-05, March 2012.

[UNIQUE] Zhang, M., Eastlake, D., Perlman, R., Cullen, M., Zhai, H., and D. Liu, "TRILL Multilevel Using Unique Nicknames", Work in Progress, draft-ietf-trill-multilevel-unique-nickname-02, May 2017.

[SingleName] Zhang, M., Eastlake, D., Perlman, R., Cullen, M., and H. Zhai, "Transparent Interconnection of Lots of Links (TRILL) Single Area Border RBridge Nickname for Multilevel", Work in Progress, draft-ietf-trill-multilevel-single-nickname-04, July 2017.

Acknowledgements

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

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

Authors' Addresses

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

Email: radia@alum.mit.edu

Donald Eastlake 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757
United States of America

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

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

Email: zhangmingui@huawei.com

Anoop Ghanwani
Dell
5450 Great America Parkway
Santa Clara, CA 95054
United States of America

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