INTERNET-DRAFT
Intended Status: Proposed Standard
Updates: 6325

Mingui Zhang
Huawei
Tissa Senevirathne
Consultant
Janardhanan Pathangi
Gigamon
Ayan Banerjee
Cisco
Anoop Ghanwani
DELL

Expires: July 23, 2018
January 19, 2018

TRILL: Resilient Distribution Trees
draft-ietf-trill-resilient-trees-09.txt

Abstract

The TRILL (Transparent Interconnection of Lots of Links) protocol provides multicast data forwarding based on IS-IS link state routing. Distribution trees are computed based on the link state information through Shortest Path First calculation. When a link on the distribution tree fails, a campus-wide re-convergence of this distribution tree will take place, which can be time consuming and may cause considerable disruption to the ongoing multicast service.

This document specifies how to build backup distribution trees to protect links on the primary distribution tree. Since the backup distribution tree is built up ahead of the link failure, when a link on the primary distribution tree fails, the pre-installed backup forwarding table will be utilized to deliver multicast packets without waiting for the campus-wide re-convergence. This minimizes the service disruption. This document updates RFC 6325.

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at

http://www.ietf.org/1id-abstracts.html

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

Table of Contents

1. Introduction

Lots of multicast traffic is generated by interrupt latency sensitive applications, e.g., video distribution including IPTV, video conference and so on. Normally, a network fault will be recovered through a network wide re-convergence of the forwarding states, but this process is too slow to meet tight Service Level Agreement (SLA) requirements on the duration of service disruption.

Protection mechanisms are commonly used to reduce the service disruption caused by network faults. With backup forwarding states installed in advance, a protection mechanism can restore an interrupted multicast stream in a much shorter time than the normal network wide re-convergence, which can meet stringent SLAs on service disruption. A protection mechanism for multicast traffic has been developed for IP/MPLS networks [RFC7431]. However, TRILL constructs distribution trees (DT) in a different way from IP/MPLS; therefore a multicast protection mechanism suitable for TRILL is developed in this document.

This document specifies "Resilient Distribution Trees" in which backup trees are installed in advance for the purpose of fast failure repair. Three types of protection mechanisms are specified.

- Global 1:1 protection refers to the mechanism where the multicast source RBridge normally injects one multicast stream onto the primary DT. When an interruption of this stream is detected, the source RBridge switches to the backup DT to inject subsequent multicast streams until the primary DT is recovered.

- Global 1+1 protection refers to the mechanism where the multicast source RBridge always injects two copies of multicast streams, one onto the primary DT and one onto the backup DT respectively. In the normal case, multicast receivers pick the stream sent along the primary DT and egress it to its local link. When a link failure interrupts the primary stream, the backup stream will be picked until the primary DT is recovered.

- Local protection refers to the mechanism where the RBridge attached to the failed link locally repairs the failure.

Resilient Distribution Trees can greatly reduce the service disruption caused by link failures. In the global 1:1 protection, the time cost for DT recalculation and installation can be saved. The global 1+1 protection and local protection further saves the time spent on the propagation of failure indication. Routing can be repaired for a failed link in tens of milliseconds.

Protection mechanisms to handle node failures are out the scope of this document. Although it's possible to use Resilient Distribution Trees to achieve load balancing of multicast traffic, this document leaves that for future study.

[RFC7176] specifies the Affinity Sub-TLV. An "Affinity Link" can be explicitly assigned to a distribution tree or trees as discussed in Section 2.1. This offers a way to manipulate the calculation of distribution trees. With intentional assignment of Affinity Links, a backup distribution tree can be set up to protect links on a primary distribution tree.

This document updates [RFC6325] as specified in Section 5.3.1.

1.1. Conventions used in this document

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

1.2. Terminology

BFD: Bidirectional Forwarding Detection [RFC7175] [RBmBFD]

CMT: Coordinated Multicast Trees [RFC7783]

Child: A directly connected node further from the Root.

DT: Distribution Tree [RFC6325]

IS-IS: Intermediate System to Intermediate System [RFC7176]

LSP: IS-IS Link State PDU

mLDP: Multipoint Label Distribution Protocol [RFC6388]

MPLS: Multi-Protocol Label Switching

Parent: A directly connected node closer to the Root.

PDU: Protocol Data Unit

Root: The top node in a tree.

PIM: Protocol Independent Multicast [RFC7761]

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

local protection (Section 5.3).

RBridge: A device implementing the TRILL protocol [RFC6325] [RFC7780]

RPF: Reverse Path Forwarding

SLA: Service Level Agreement

Td: failure detection timer

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

2. Usage of the Affinity Sub-TLV

This document uses the already existing Affinity Sub-TLV [RFC7176] to assign a parent to an RBridge in a tree as discussed below. Support of the Affinity Sub-TLV by an RBridge is indicated by a capability bit in the TRILL-VER Sub-TLV [RFC7783].

2.1. Indicating Affinity Links

The Affinity Sub-TLV explicitly assigns parents for RBridges on distribution trees. It is distributed in an LSP and can be recognized by each RBridge in the campus. The originating RBridge becomes the parent and the nickname contained in the Affinity Record identifies the child. This explicitly provides an "Affinity Link" on a distribution tree or trees. The "Tree-num of roots" in the Affinity Record(s) in the Affinity Sub-TLV identify the distribution trees that adopt this Affinity Link [RFC7176].

Suppose the link between RBridge RB2 and RBridge RB3 is chosen as an Affinity Link on the distribution tree rooted at RB1 in Figure 2.1. RB2 sends out the Affinity Sub-TLV with an Affinity Record that says {Nickname=RB3, Num of Trees=1, Tree-num of roots=RB1}. Different from the Affinity Link usage in [RFC7783], RB3 does not have to be a leaf node on a distribution tree. Therefore an Affinity Link can be used to identify any link on a distribution tree. This kind of assignment offers a flexibility of control to RBridges in distribution tree calculation: they can be directed to choose a child for which they are not on the shortest paths from the root. This flexibility is used to construct back-up trees that can be used to increase the reliability of distribution trees. Affinity Links may be configured or automatically determined according to an algorithm as described in this document.

Affinity Link SHOULD NOT be misused to declare connection of two RBridges that are not adjacent. If it is, the Affinity Link is

ignored and has no effect on tree building.

2.2. Distribution Tree Calculation with Affinity Links

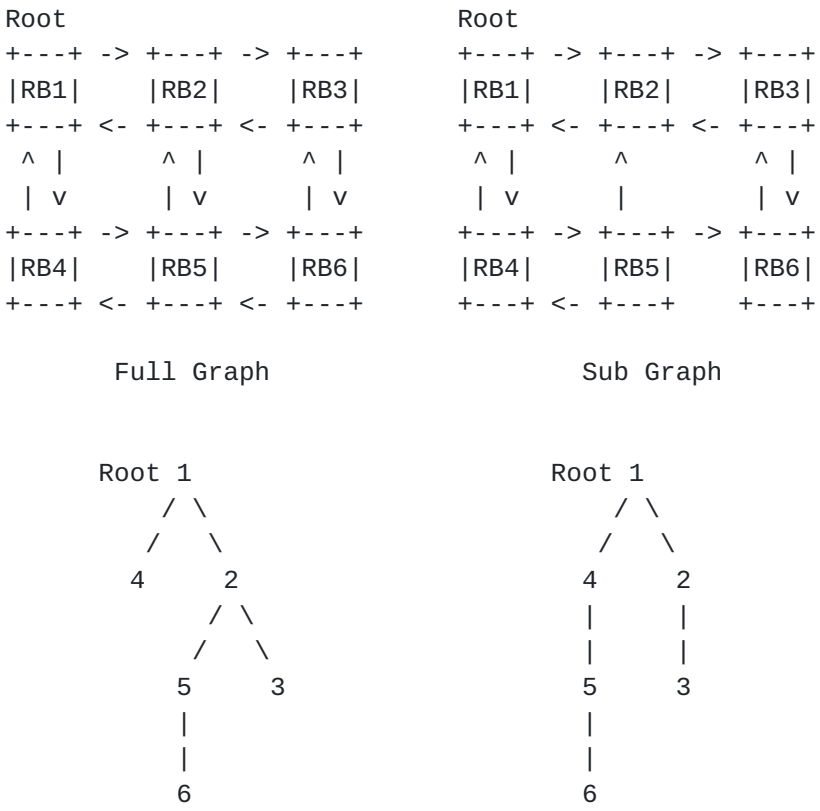


Figure 2.1: DT Calculation with the Affinity Link RB4-RB5

When RBridges receive an Affinity Sub-TLV declaring an Affinity Link that is an incoming link of an RBridge (i.e., this RBridge is the child on this Affinity Link) for a particular distribution tree, this RBridge's incoming links/adjacencies other than the Affinity Link are removed from the full graph of the campus to get a sub graph to compute that tree. RBridges perform the Shortest Path First calculation to compute the tree based on the resulting sub graph. This assures that the Affinity Link appears in the distribution tree being calculated.

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

Note that outgoing links/adjacencies are not affected by the Affinity Link. When two RBridges, say RB4 and RB5, are adjacent, the adjacency/link from RB4 to RB5 and the adjacency/link from RB5 to RB4 are separate and, for example, might have different costs.

3. Distribution Tree Calculation

RBridges use IS-IS to advertise adjacencies and thus advertise network faults through the withdrawal of such adjacencies. A node or link failure will trigger a campus-wide re-convergence of all TRILL distribution trees. The re-convergence generally includes the following sequence of procedures:

1. Failure (loss of adjacency) detected through IS-IS control messages (HELLO) not getting through or some other link test such as BFD [RFC7175] [RBmBFD];

2. IS-IS state flooding so each RBridge learns about the failure;

3. Each RBridge recalculates affected distribution trees independently;

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

The re-convergence time to go through these four steps disrupts ongoing multicast traffic. In protection mechanisms, alternative paths prepared ahead of potential node or link failures are available to detour around the failures upon the failure detection; thus service disruption can be minimized.

This document focuses only on link failure protection. The construction of backup DTs (distribution trees) for the purpose of node protection is out of scope. (The usual way to protect from a node failure on the primary tree, is to have a backup tree setup without this node. When this node fails, the backup tree can be safely used to forward multicast traffic to make a detour. However, TRILL distribution trees are shared among all VLANs and Fine Grained Labels [RFC7172] and they have to cover all RBridge nodes in the campus [RFC6325]. A DT that does not span all RBridges in the campus may not cover all receivers of many multicast groups. (This is different from the multicast trees construction signaled by PIM (protocol independent multicast [RFC7761]) or mLDP (multicast label distribution protocol [RFC6388].))

3.1. Designating Roots for Backup Distribution Trees

The RBridge, say, RB1, having the highest root priority nickname controls the creation of backup DTs and specifies their roots. It explicitly advertises a list of nicknames identifying the roots of primary and their backup DTs using the Backup Tree APPsub-TLV as specified in Section 6.2 (See also Section 4.5 of [RFC6325]). It's possible that a backup DT and a primary DT have the same root RBridge but this is not required. In that case, to distinguish the primary DT and the backup DT for the common root case, the root RBridge MUST own at least two nicknames so a different nickname can be used to name each tree.

The method by which the highest priority root RBridge determines which primary distribution trees to protect with a backup and what the root of each such back up will be is out of scope for this document.

3.2. Backup DT Calculation with Affinity Links

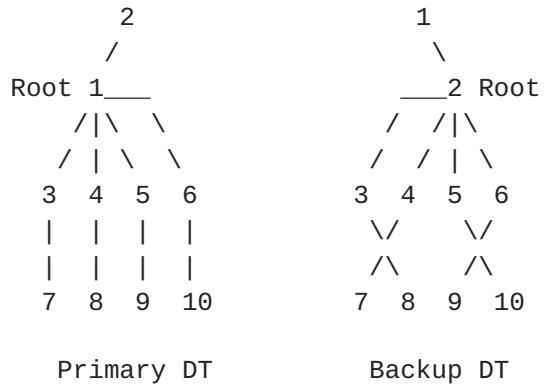


Figure 3.1: An Example of a Primary DT and its Backup DT

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

3.2.1. The Algorithm for Choosing Affinity Links

Operators MAY configure Affinity Links, for example, to intentionally protect a specific link such as the link connected to a gateway. But it is desirable that every RBridge independently computes Affinity

Links for a backup DT across the whole campus. This enables a distributed deployment and also minimizes configuration.

Compared to the algorithms for Maximally Redundant Trees in [RFC7811], TRILL has both an advantage and a disadvantage. An advantage of TRILL is that Resilient Distribution Tree does not restrict the root of the backup DT to be the same as that of the primary DT. Two disjoint (or maximally disjoint) trees may have different root nodes, which significantly augments the solution space.

A disadvantage of TRILL, when using the algorithm specified below in this section is that the backup DT is computed with reference to the primary tree but there may be a pair of tree that is more disjoint than any backup tree can be with the particular primary tree.

This document RECOMMENDS achieving the independent backup tree determination method through a change to the conventional DT calculation process of TRILL. After the primary DT is calculated, every RBridge will be aware of which links are used in that primary tree. When the backup DT is calculated, each RBridge increases the metric of these links by the summation of all original link metrics in the campus but not more than 2**23, which gives these links a lower priority of being chosen for the backup DT by the Shortest Path First calculation. All links on this backup DT can be assigned as Affinity Links but this may not be necessary. In order to reduce the amount of Affinity Sub-TLVs flooded across the campus, only those NOT picked by the conventional DT calculation process SHOULD be announced as Affinity Links.

3.2.2. Affinity Links Advertisement

Similar to [RFC7783], every parent RBridge of an Affinity Link takes charge of announcing this link in an Affinity Sub-TLV. When this RBridge plays the role of parent RBridge for several Affinity Links, it is natural to have them advertised together in the same Affinity Sub-TLV, and each Affinity Link is structured as one Affinity Record [RFC7176].

Affinity Links are announced in the Affinity Sub-TLV that is recognized by every RBridge. Since each RBridge computes distribution trees as the Affinity Sub-TLV requires, the backup DT will be built consistently by all RBridges in the campus.

4. Resilient Distribution Trees Installation

As specified in Section 4.5.2 of [RFC6325], an ingress RBridge MUST announce the distribution trees it may choose to ingress multicast

frames. Thus other RBridges in the campus can limit the amount of state necessary for RPF checks. Also, [RFC6325] recommends that an ingress RBridge by default chooses the DT or DTs whose root or roots are least cost from the ingress RBridge. To sum up, RBridges do pre-compute all the trees that might be used so they can properly forward multi-destination packets, but only install RPF state for some combinations of ingress and tree.

This document specifies that the backup DT MUST be included in an ingress RBridge's DT announcement list in this ingress RBridge's LSP if the corresponding primary tree is included. In order to reduce the service disruption time, RBridges SHOULD install backup DTs in advance, which also includes the RPF filters that need to be set up for RPF Checks.

Since the backup DT is intentionally built highly disjoint to the primary DT, when a link fails and interrupts the ongoing multicast traffic sent along the primary DT, it is probable that the backup DT is not affected. Therefore, the backup DT installed in advance can be used to deliver multicast packets immediately.

4.1. Pruning the Backup Distribution Tree

The way that a backup DT is pruned is different from the way that the primary DT is pruned. To enable protection it is possible that a branch should not be pruned (see Section 4.5.3 of [RFC6325]), even though it does not have any downstream receivers for a particular data label. The rule for backup DT pruning is that the backup DT should be pruned, eliminating branches that have no potential downstream RBridges which appear on the pruned primary DT.

Even though the primary DT may not be optimally pruned in practice, the backup DT SHOULD always be pruned as if the primary DT is optimally pruned. Those redundant links that ought to be pruned on the primary DT will not be protected.

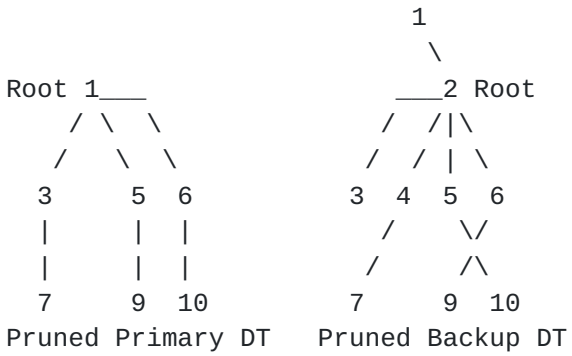


Figure 4.1: The Backup DT is Pruned Based on the Pruned Primary DT.

Suppose RB7, RB9 and RB10 constitute a multicast group MGx. The pruned primary DT and backup DT are shown in Figure 4.1. Referring back to Figure 3.1, branches RB2-RB1 and RB4-RB1 on the primary DT are pruned for the distribution of MGx traffic since there are no potential receivers on these two branches. Although branches RB1-RB2 and RB3-RB2 on the backup DT have no potential multicast receivers, they appear on the pruned primary DT and may be used to repair link failures of the primary DT. Therefore they are not pruned from the backup DT. Branch RB8-RB3 can be safely pruned because it does not appear on the pruned primary DT.

4.2. RPF Filters Preparation

RB2 announces in its LSP the trees RB2 might choose when RB2 ingresses a multicast packet [RFC6325]. When RB2 specifies such trees, it SHOULD include the backup DT. Other RBridges will prepare the RPF check states for both the primary DT and backup DT. When a multicast packet is sent along either the primary DT or the backup DT, it will be subject to the RPF Check. This works when global 1:1 protection is used. However, when global 1+1 protection or local protection is applied, traffic duplication will happen if multicast receivers accept both copies of the multicast packets from two RPF filters. In order to avoid such duplication, egress RBridge multicast receivers MUST act as merge points to activate a single RPF filter and discard the duplicated packets from the other RPF filter. In the normal case, the RPF state is set up according to the primary DT. When a link failure on the primary DT is detected, the egress node RPF filter based on the backup DT should be activated.

5. Protection Mechanisms with Resilient Distribution Trees

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

- The link on a TRILL distribution tree is always bidirectional while the link on a distribution tree in IP/MPLS networks may be unidirectional.

- In TRILL, a multicast source node does not have to be the root of the distribution tree. It is just the opposite in IP/MPLS networks.

- In IP/MPLS networks, distribution trees are constructed for each multicast source node as well as their backup distribution trees. In TRILL, a small number of core distribution trees are shared among multicast groups. A backup DT does not have to share the

same root as the primary DT.

Therefore a TRILL specific multicast protection mechanism is needed.

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

The type of protection in use at an RBridge is indicated by a two-bit field in that RBridge's Extended Capability TLV as discussed in Section 5.4.

5.1. Global 1:1 Protection

In the global 1:1 protection, the ingress RBridge of the multicast traffic is responsible for switching the failure affected traffic from the primary DT over to the backup DT. Since the backup DT has been installed in advance, the global protection need not wait for the DT recalculation and installation. When the ingress RBridge is notified about the failure, it immediately makes this switch over.

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

BFD (Bidirectional Forwarding Detection) protocol can be used to reduce the failure detection time. Link failures can be rapidly detected with one-hop BFD [RFC7175]. [RBmBFD] introduces the fast failure detection of multicast paths. It can be used to reduce both the failure detection and the propagation time for global protection. In [RBmBFD], the ingress RBridge needs to send BFD control packets to poll each receiver, and receivers return BFD control packets to the ingress as the response. If no response is received from a specific receiver for a detection time, the ingress can judge that the connectivity to this receiver is broken. Therefore, [RBmBFD] is used to detect the connectivity of a path rather than a link. The ingress RBridge will determine a minimum failed branch that contains this receiver. The ingress RBridge will switch ongoing multicast traffic

based on this judgment. For example, in Figure 4.1, if RB9 does not respond while RB10 still responds, RB7 will presume that link RB1-RB5 and RB5-RB9 are failed. Multicast traffic will be switched to a backup DT that can protect these two links. More accurate link failure detection might help ingress RBridges make smarter decision but it's out of the scope of this document.

5.2. Global 1+1 Protection

In the global 1+1 protection, the multicast source RBridge always replicates the multicast packets and sends them onto both the primary and backup DT. This may sacrifice the capacity efficiency but given there is much connection redundancy and inexpensive bandwidth in Data Center Networks, such kind of protection can be popular [RFC7431].

5.2.1. Failure Detection

Egress RBridges (merge points) SHOULD realize the link failure as early as practical and update their RPF filters quickly to minimize the traffic disruption. Three options are provided as follows.

1. If you had a very reliable and steady data stream, egress RBridges assume a minimum known packet rate for that data stream [RFC7431]. A failure detection timer (say Td) is set as the interval between two continuous packets. Td is reinitialized each time a packet is received. If Td expires and packets are arriving at the egress RBridge on the backup DT (within the time frame Td), it updates the RPF filters and starts to receive packets forwarded on the backup DT. This method requires configuration at the egress RBridge of Td and of some method (filter) to determine if a packet is part of the reliable data stream. Since the filtering capabilities of various fast path logic differs greatly, specifics of such configuration are outside the scope of this document.

2. With multi-point BFD [RBmBFD], when a link failure happens, affected egress RBridges can detect a lack of connectivity from the ingress. Therefore these egress RBridges are able to update their RPF filters promptly.

3. Egress RBridges can always rely on the IS-IS control plane to learn the failure and determine whether their RPF filters should be updated.

5.2.2. Traffic Forking and Merging

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

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

Note that the egress RBridge need not be a literal merge point, that is receiving the primary and backup DT versions over different links. Even if the egress RBridge receives both copies over the same link, because disjoint links are not available, it can still filter out one copy because the RFP filtering logic is designed to test which tree the packet is on as indicated by a field in the TRILL Header [RFC6325].

5.3. Local Protection

In the local protection, the Point of Local Repair (PLR) happens at the upstream RBridge connected to the failed link. It is this RBridge that makes the decision to replicate the multicast traffic to recover from this link failure. Local protection can further save the time spent on failure notification through the flooding of LSPs across the TRILL campus. In addition, the failure detection can be sped up using BFD [RFC7175], therefore local protection can minimize the service disruption, typically reducing it to less than 50 milliseconds.

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

Due to the multi-destination RPF check in TRILL, local protection can only be used at a fork point where the primary and backup trees diverge and the set of nodes downstream is identical for both paths. If these conditions do not apply, local protection MUST NOT be used.

5.3.1. Starting to Use the Backup Distribution Tree

The egress nickname TRILL Header field of the replicated multicast TRILL data packets specifies the tree on which they are being distributed. This field will be rewritten to the backup DT's root nickname by the PLR. But the ingress nickname field of the multicast TRILL Data packet MUST remain unchanged. The PLR forwards all multicast traffic with the backup DT egress nickname along the backup DT. This updates [RFC6325] which specifies that the egress nickname in the TRILL header of a multi-destination TRILL data packet must not be changed by transit RBridges.

In the above example, the PLR RB1 locally decides to send replicated multicast packets according to the backup DT. It will send them to the next hop RB2.

5.3.2. Duplication Suppression

When a PLR starts to send replicated multicast packets on the backup DT, some multicast packets are still being sent along the primary DT. Some egress RBridges might receive duplicated multicast packets. The traffic forking and merging method in the global 1+1 protection can be adopted to suppress the duplication.

5.3.3. An Example to Walk Through

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

In the normal case, multicast frames ingressed by RB7 in Figure 4.1 with pruned distribution on the primary DT rooted at RB1 are being received by RB9 and RB10. When the link RB1-RB5 fails, the PLR RB1 begins to replicate and forward subsequent multicast packets using the pruned backup DT rooted at RB2. When RB2 gets the multicast packets from the link RB1-RB2, it accepts them since the RPF filter {DT=RB2, ingress=RB7, receiving links=RB1-RB2, RB3-RB2, RB4-RB2, RB5-RB2 and RB6-RB2} is installed on RB2. RB2 forwards the replicated multicast packets to its neighbors except RB1. The multicast packets reach RB6 where both RPF filters {DT=RB1, ingress=RB7, receiving link=RB1-RB6} and {DT=RB2, ingress=RB7, receiving links=RB2-RB6 and RB9-RB6} are active. RB6 will let both multicast streams through. Multicast packets will finally reach RB9 where the RPF filter is updated from {DT=RB1, ingress=RB7, receiving link=RB5-RB9} to {DT=RB2, ingress=RB7, receiving link=RB6-RB9}. RB9 will egress the multicast packets from the Backup Distribution Tree on to the local link and drop those from the Primary Distribution Tree based on the reverse path forwarding filter.

5.4. Protection Mode Signaling

The desired mode of resilient tree operation for each RBridge is chosen by the network operator and configured on that RBridge. This mode is announced by each RBridge is a two-bit Resilient Tree Mode field in their Extended Capabilities TLV (see Sections 6.1, 8.1). The values of this field have the following meanings:

Value	Short Name	Effect
00	No support	If any RBridge does not support Resilient Trees, then the Resilient Tree mechanism is

		disabled in all RBridges. This also applies if any RBridge does not announce an Extended Capabilities TLV.
01	Global 1:1	An RBridge advertising this value will, when it ingresses a multi-destination frames, send them on only one of the primary and backup DTs. All other RBridges set their RPF filters to accept traffic on both trees from this ingress.
10	Global 1+1	An RBridge advertising this value will, when it ingresses a multi-destination frames, send them on both the primary and backup DTs. All other RBridges MUST set their RPF filters to accept Traffic only on the primary or backup DT.
11	1+1 & Local	An RBridge advertising this value acts as an for the value 01 above when it is the ingress RBridge. In addition, if it is a transit RBridge at a fork point between the primary and backup tress and detects that an adjacency has failed, it diverts multi-destination TRILL data packts on the primary tree to the backup tree, changing the tree id in the packet to the backup tree.

5.5. Updating the Primary and the Backup Distribution Trees

Assume an RBridge receives the LSP that indicates a link failure. This RBridge starts to calculate the new primary DT based on the new topology with the failed link excluded. Suppose the new primary DT is installed at t1.

The propagation of LSPs around the campus will take some time. For safety, we assume all RBridges in the campus will have converged to the new primary DT at t1+Ts. By default, Ts (the "settling time") is set to 30 seconds but it is configurable in seconds from 1 to 100. At t1+Ts, the ingress RBridge switches the traffic from the backup DT back to the new primary DT.

After another Ts (at t1+2*Ts), no multicast packets are being forwarded along the old primary DT. The backup DT should be updated (recalculated and reinstalled) after the new primary DT. The process of this update under different protection types are discussed as follows.

a) For the global 1:1 protection, the backup DT is simply updated at t1+2*Ts.

b) For the global 1+1 protection, the ingress RBridge stops replicating the multicast packets onto the old backup DT at t1+Ts.

The backup DT is updated at t1+2*Ts. The ingress RBridge MUST wait for another Ts, during which time period all RBridges converge to the new backup DT. At t1+3*Ts, it's safe for the ingress RBridge to start to replicate multicast packets onto the new backup DT.

c) For the local protection, the PLR stops replicating and sending packets on the old backup DT at t1+Ts. It is safe for RBridges to start updating the backup DT at t1+2*Ts.

6. TRILL IS-IS Extensions

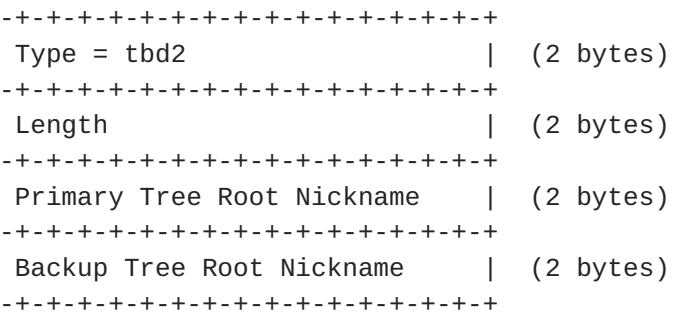
This section lists extensions to TRILL IS-IS to support resilient trees.

6.1. Resilient Trees Extended Capability Field

An RBridge that supports the facilities specified in this document MUST announce the Extended RBridge Capabilities APPsub-TLV [RFC7782] with a non-zero value in the Resilient Trees field. If there are RBridges that do not announce field set to a non-zero value, all RBridges of the campus MUST disable the Resilient Distribution Tree mechanism as defined in this document and fall back to the distribution tree calculation algorithm as specified in [RFC6325].

6.2 Backup Tree Root APPsub-TLV

The structure of the Backup Tree Root APPsub-TLV is shown below.



- Type = Backup Tree Root APPsubTLV type, set to tbd2
- Length = 4, if the length is any other value, the APPsub-TLV is corrupt and MUST be ignored.
- Primary Tree Root Nickname = the nickname of the root RBridge of the primary tree for which a resilient backup tree is being created
- Backup Tree Root Nickname = the nickname of the root RBridge of

the backup tree

If either nickname is not the nickname of a tree whose calculation is being directed by the highest priority tree root RBridge, the APPsub-TLV is ignored. This APPsub-TLV MUST be advertised by the highest priority RBridge to be a tree root. Backup Tree Root APPsub-TLVs advertised by other RBridges are ignored. If there are two or more Backup Tree Root APPsub-TLVs for the same primary tree specifying different backup trees, then the one specifying the lowest magnitude backup tree root nickname is used, treating nicknames as unsigned 16-bit quantities.

7. Security Considerations

This document raises no new security issues for TRILL. The IS-IS PDUs used to transmit the information specified in Section 6 can be secured with IS-IS security [RFC5310].

For general TRILL Security Considerations, see [RFC6325].

8. IANA Considerations

The Affinity Sub-TLV has already been defined in [RFC7176]. This document does not change its definition. See below for IANA Actions.

8.1. Resilient Tree Extended Capability Field

IANA will assign two adjacent bits (Sections 5.4, 6.1) in the Extended RBridge Capabilities subregistry on the TRILL Parameters page to form the Resilient Tree Extended Capability field and change the heading of the "Bit" column to be "Bit(s)", adding the following to the registry [for example, tbd1 could be "2-3"]:

Bit	Mnemonic	Description	Reference
tbd1	RT	Resilient Tree Support	[this document]

8.2. Backup Tree Root APPsub-TLV

IANA will assign and APPsub-TLV type under IS-IS TLV 251 Application Identifier 1 on the TRILL Parameters page from the range below 255 for the Backup Tree Root APPsub-TLV (Section 6.2) as follows:

Type	Name	Reference
tbd2	Backup Tree Root	[this document]

Acknowledgements

The careful review from Gayle Noble is gracefully acknowledged. The authors would like to thank the comments and suggestions from Donald Eastlake, Erik Nordmark, Fangwei Hu, Gayle Noble, Hongjun Zhai and Xudong Zhang.

9. References

9.1. Normative References

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

[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, <http://www.rfc-editor.org/info/rfc7176>.

[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>.

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

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <http://www.rfc-editor.org/info/rfc7761>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <http://www.rfc-editor.org/info/rfc6388>.

[RBmBFD] M. Zhang, S. Pallagatti and V. Govindan, "TRILL Support of Point to Multipoint BFD", draft-ietf-trill-p2mp-bfd, work in progress.

[RFC7175] Manral, V., Eastlake 3rd, D., Ward, D., and A. Banerjee,

"Transparent Interconnection of Lots of Links (TRILL): Bidirectional Forwarding Detection (BFD) Support", RFC 7175, DOI 10.17487/RFC7175, May 2014, <http://www.rfc-editor.org/info/rfc7175>.

[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>.

[RFC7782] Zhang, M., Perlman, R., Zhai, H., Durrani, M., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL) Active-Active Edge Using Multiple MAC Attachments", RFC 7782, DOI 10.17487/RFC7782, February 2016, <http://www.rfc-editor.org/info/rfc7782>.

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

9.2. Informative References

[RFC7811] Enyedi, G., Csaszar, A., Atlas, A., Bowers, C., and A. Gopalan, "An Algorithm for Computing IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7811, DOI 10.17487/RFC7811, June 2016, <http://www.rfc-editor.org/info/rfc7811>.

[RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <http://www.rfc-editor.org/info/rfc7431>.

[mBFD] D. Katz, D. Ward, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint, work in progress.

[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, <http://www.rfc-editor.org/info/rfc7172>.

Author's Addresses

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

Email: zhangmingui@huawei.com

Tissa Senevirathne
Consultant

Email: tsenevir@gmail.com

Janardhanan Pathangi
Gigamon

Email: path.jana@gmail.com

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

Email: ayabaner@cisco.com

Anoop Ghanwani
Dell
350 Holger Way
San Jose, CA 95134

Phone: +1-408-571-3500
Email: Anoop@alumni.duke.edu