INTERNET-DRAFT                                           Mingui Zhang
Intended Status: Proposed Standard                            Huawei
Expires: April 24, 2014                          Tissa Senevirathne
Updates: RFC 6325                                              CISCO
                                               Janardhanan Pathangi
                                                               DELL
                                                      Ayan Banerjee
                                                   Insieme Networks
                                                    Anoop Ghanwani
                                                               DELL
                                                   Donald Eastlake
                                                             Huawei
                                                   October 21, 2013

                    TRILL Resilient Distribution Trees
                  draft-zhang-trill-resilient-trees-04.txt

Abstract

   TRILL protocol provides layer 2 multicast data forwarding using 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 recovergence of
   this distribution tree will take place, which can be time consuming
   and may cause considerable disruption to the ongoing multicast
   service.

   This document proposes to build the backup distribution tree 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 recovergence, which minimizes the
   service disruption.

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


Copyright and License Notice

Table of Contents

1. Introduction

   Lots of multicast traffic is generated by interrupt latency sensitive
   applications, e.g., video distribution, including IP-TV, video
   conference and so on. Normally, a network fault will be recovered
   through a network wide reconvergence of the forwarding states, but
   this process is too slow to meet the tight Service Level Agreement
   (SLA) requirements on the service disruption duration. What is worse,
   updating multicast forwarding states may take significantly longer
   than unicast convergence since multicast states are updated based on
   control-plane signaling [mMRT].

   Protection mechanisms are commonly used to reduce the service
   disruption caused by network faults. With backup forwarding states
   installed in advance, a protection mechanism is possible to restore
   an interrupted multicast stream in tens of milliseconds which
   guarantees the stringent SLA on service disruption. Several
   protection mechanisms for multicast traffic have been developed for
   IP/MPLS networks [mMRT] [MoFRR]. However, the way TRILL constructs
   distribution trees (DT) is different from the way multicast trees are
   computed under IP/MPLS, therefore a multicast protection mechanism
   suitable for TRILL is required.

   This document proposes "Resilient Distribution Trees" (RDT) in which
   backup trees are installed in advance for the purpose of fast failure
   repair. Three types of protection mechanisms are proposed.

   o  Global 1:1 protection is used to refer to the mechanism that the
      multicast source RBridge normally injects one multicast stream
      onto the primary DT. When 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.

   o  Global 1+1 protection is used to refer to the mechanism that the
      multicast source RBridge always injects two copies of multicast
      streams onto the primary DT and 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 one will be picked until
      the primary DT is recovered.

   o  Local protection refers to the mechanism that the RBridge attached
      to the failed link locally repairs the failure.

   RDT may greatly reduce the service disruption caused by link
   failures. In the global 1:1 protection, the time cost by DT
   recalculation and installation can be saved. The global 1+1
   protection and local protection further save the time spent on

failure propagation. A failed link can be repaired in tens of
milliseconds. Although it's possible to make use of RDT to achieve
load balance of multicast traffic, this document leaves that for
future study.

[6326bis] defines the Affinity TLV. An "Affinity Link" can be
explicitly assigned to a distribution tree or trees. 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.

## 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

IS-IS: Intermediate System to Intermediate System
TRILL: TRansparent Interconnection of Lots of Links
DT: Distribution Tree
RPF: Reverse Path Forwarding
RDT: Resilient Distribution Tree
SLA: Service Level Agreement
PLR: Point of Local Repair, in this document, it is the multicast
  upstream RBridge connecting the failed link. It's valid only for
  local protection.

## 2. Usage of Affinity Sub-TLV

This document uses the Affinity Sub-TLV [6326bis] to assign a parent
to an RBridge in a tree as discussed below.

## 2.1. Allocating Affinity Links

Affinity Sub-TLV explicitly assigns parents for RBridges on
distribution trees. They are advertised in the Affinity Sub-TLV and
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" of the Affinity
Record identify the distribution trees that adopt this Affinity Link
[6326bis].

Affinity Links may be configured or automatically determined using a
certain algorithm [CMT]. Suppose link RB2-RB3 is chosen as an
Affinity Link on the distribution tree rooted at RB1. RB2 should send

out the Affinity Sub-TLV with an Affinity Record like {Nickname=RB3, Num of Trees=1, Tree-num of roots=RB1}. In this document, 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 to RBridges in distribution tree calculation: they are allowed to choose child for which they are not on the shortest paths from the root. This flexibility is leveraged to increase the reliability of distribution trees in this document.

An Affinity Sub-TLV which tries to connect two RBridges that are not adjacent MUST be ignored.

2.2. Distribution Tree Calculation with Affinity Links

When RBridges receive an Affinity Sub-TLV with Affinity Link which is an incoming link of RB2 (i.e., RB2 is the child on this Affinity Link), RB2's incoming links other than the Affinity Link are removed from the full graph of the campus to get a sub graph. RBridges perform Shortest Path First calculation to compute the distribution tree based on the sub graph. In this way, the Affinity Link will surely appear on the distribution tree.

```
         Root                              Root
         +---+ -> +---+ -> +---+           +---+ -> +---+ -> +---+
         |RB1|    |RB2|    |RB3|           |RB1|    |RB2|    |RB3|
         +---+ <- +---+ <- +---+           +---+ <- +---+ <- +---+
          ^ |      ^ |      ^ |             ^ |      ^        ^ |
          | v      | v      | v             | v      |        | v
         +---+ -> +---+ -> +---+           +---+ -> +---+ -> +---+
         |RB4|    |RB5|    |RB6|           |RB4|    |RB5|    |RB6|
         +---+ <- +---+ <- +---+           +---+ <- +---+    +---+

              Full Graph                        Sub Graph


            Root 1                            Root 1
             / \                               / \
            /   \                             /   \
           4     2                           4     2
                / \                          |     |
               /   \                         |     |
              5     3                        5     3
              |                              |
              |                              |
              6                              6
```
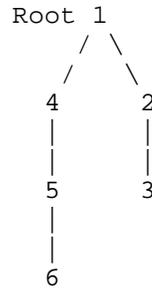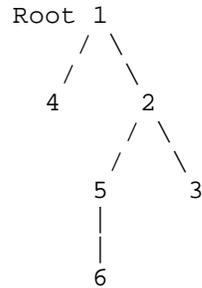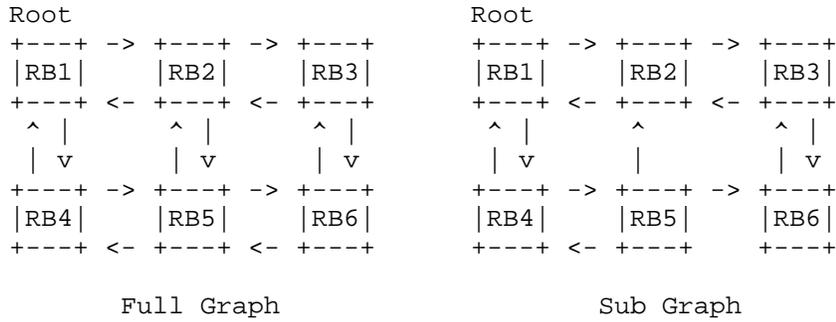
    Shortest Path Tree of Full Graph   Shortest Path Tree of Sub Graph

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

    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.

3. Resilient Distribution Trees Calculation

    RBridges leverage IS-IS to detect and advertise network faults. A
    node or link failure will trigger a campus-wide reconvergence of
    distribution trees. The reconvergence generally includes the
    following procedures:

    1. Failure detected through IS-IS control messages (HELLO) exchanging
       or some other method such as BFD [rbBFD];

    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 per VLAN and
      installed into the multicast forwarding tables.

   The slow reconvergence can be as long as tens of seconds, which will
   cause disruption to ongoing multicast traffic. In protection
   mechanisms, alternative paths prepared ahead of potential node or
   link failures are used to detour the failures upon the failure
   detection, therefore service disruption can be minimized.

   This document will focus only on link protection. The construction of
   backup DT for the purpose of node protection is out the scope of this
   document. In order to protect a node on the primary tree, a backup
   tree can be setup without this node [mMRT]. 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 [FGL] 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 [RFC4601] or mLDP [RFC6388].)

3.1. Designating Roots for Backup Trees

   Operators MAY manually configure the roots for the backup DTs.
   Nevertheless, this document aims to provide a mechanism with minimum
   configuration. Two options are offered as follows.

3.1.1. Conjugate Trees

   [RFC6325] and [ClearC] has defined how distribution tree roots are
   selected. When a backup DT is computed for a primary DT, its root is
   set to be the root of this primary DT. In order to distinguish the
   primary DT and the backup DT, the root RBridge MUST own multiple
   nicknames.

3.1.2. Explicitly Advertising Tree Roots

   RBridge RB1 having the highest root priority nickname might
   explicitly advertise a list of nicknames to identify the roots of the
   primary and backup tree roots (See [RFC6325] Section 4.5).

3.2. Backup DT Calculation

3.2.1. Backup DT Calculation with Affinity Links

```
                2                        1
               /  \                     \
        Root 1___               ___2 Root
          /|\  \                  /  /|\
         / | \  \                /  / | \
        3  4  5  6             3  4  5  6
        |  |  |  |              \/      \/
        |  |  |  |              /\      /\
        7  8  9  10            7  8  9  10

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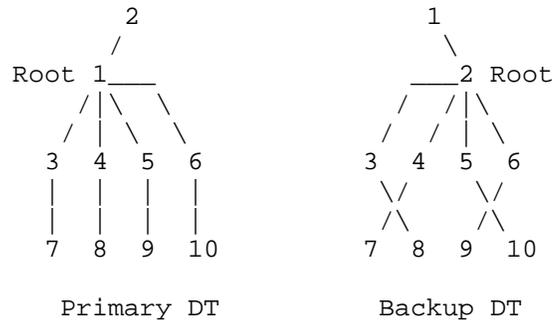
Figure 3.1: An Example of a Primary DT and its Backup DT

TRILL allows RBridges to compute multiple distribution trees. With
the intentional assignment of Affinity Links in DT calculation, this
document proposes a method to construct Resilient Distribution Trees
(RDT). For example, in Figure 3.1, the backup DT is set up 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 links on the backup DT. It means that every link on the
primary DT, except link RB1-RB2, can be protected by the backup DT.

3.2.1.1. Algorithm for Choosing Affinity Links

Operators MAY configure Affinity Links 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.

Algorithms for Maximally Redundant Trees [mMRT] may be used to figure
out Affinity Links on a backup DT which is maximally disjointed to
the primary DT but it only provides a subset of all possible
solutions, i.e., the conjugate trees described in Section 3.1.1. In
TRILL, RDT does not restrict the root of the backup DT to be the same
as that of the primary DT. Two disjoint (or maximally disjointed)
trees may root from different nodes, which significantly augments the
solution space.

This document RECOMMENDS achieving the independent method through a
slight change to the conventional DT calculation process of TRILL.
Basically, after the primary DT is calculated, the RBridge will be
aware of which links will be used. When the backup DT is calculated,
each RBridge increases the metric of these links by a proper value
(for safety, it's recommended to used the summation of all original
link metrics in the campus but not more than 2**23), which gives

these links a lower priority being chosen by the backup DT by
performing Shortest Path First calculation. All links on this backup
DT can be assigned as Affinity Links but this is unnecessary. In
order to reduce the amount of Affinity Sub-TLVs flooded across the
campus, only those not picked by conventional DT calculation process
ought to be recognized as Affinity Links.

3.2.1.2. Affinity Links Advertisement

Similar to [CMT], 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.

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
up consistently.

3.2.2. Backup DT Calculation without Affinity Links

This section provides an alternative method to set up the disjointed
backup DT.

After the primary DT is calculated, each RBridge increases the cost
of those links which are already in the primary DT by a multiplier
(For safety, 64x is RECOMMENDED.). It would ensure that a link
appears in both trees if and only if there is no other way to reach
the node (i.e. the graph would become disconnected if it were pruned
of the links in the first tree.). In other words, the two trees will
be maximally disjointed.

The above algorithm is similar as that defined in Section 3.2.1.1.
All RBridges MUST agree on the same algorithm, then the backup DT can
be calculated by each RBridge consistently and configuration is
unnecessary.

4. Resilient Distribution Trees Installation

As specified in [RFC6325] Section 4.5.2, 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
states which are necessary for RPF check. 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 states that the backup DT MUST be contained in an
ingress RBridge's DT announcement list and included in this ingress
RBridge's LSP. 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 Check.

Since the backup DT is intentionally built up maximally disjointed 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 backup DT SHOULD be pruned per-VLAN. But the way a backup DT is
pruned is different from the way that the primary DT is pruned. Even
though a branch contains no downstream receivers, it is probable that
it should not be pruned for the purpose of protection. The rule for
backup DT pruning is that the backup DT should be pruned per-VLAN,
eliminating branches that have no potential downstream RBridges which
appear on the pruned primary DT.

It is probably that the primary DT is not optimally pruned in
practice. In this case, the backup DT SHOULD be pruned presuming that
the primary DT is optimally pruned. Those redundant links that ought
to be pruned will not be protected.

```
                                     1
                                      \
          Root 1___              ___2 Root
             / \   \            /  /|\
            /   \   \          /  / | \
           3    5   6         3  4  5  6
           |    |   |          /    \/
           |    |   |         /     /\
           7    9   10       7     9  10
          Pruned Primary DT   Pruned Backup DT
```
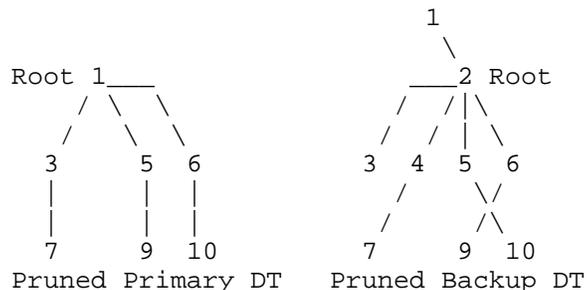
Figure 4.1: The Backup DT is Pruned Based on the Pruned Primary DT.

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 includes in its LSP the information to indicate which trees RB2
might choose to ingress multicast frames [RFC6325]. When RB2
specifies the trees it might choose to ingress multicast traffic, 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
pass 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 normal case, the RPF
state is set up according to the primary DT. When a link fails, the
RPF filter based on the backup DT should be activated.

5. Protection Mechanisms with Resilient Distribution Trees

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

o  The link on a TRILL distribution tree is bidirectional while the
   link on a distribution tree in IP/MPLS networks is unidirectional.

o  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.

o  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
developed 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 about 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 merge them. In the local
protection, when link RB1-RB5 fails, RB1 will locally replicate the
multicast traffic and send it on the backup DT.

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 considerable service disruption.

   BFD (Bidirectional Forwarding Detection) protocol can be used to
   reduce the failure detection time [rbBFD]. Link failures can be
   rapidly detected with one-hop BFD. Multi-destination BFD extends BFD
   mechanism to include the fast failure detection of multicast paths
   [mBFD]. It can be used to reduce both the failure detection and
   propagation time in the global protection. In multi-destination BFD,
   ingress RBridge need to send BFD control packets to poll each
   receiver, and receivers return BFD control packets to the ingress as
   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. In this way, multi-destination BFD detects the
   connectivity of a path rather than a link. The ingress RBridge will
   determine a minimum failed branch which contains this receiver. The
   ingress RBridge will switch ongoing multicast traffic based on this
   judgment. For example, on figure 4.1, if RB9 does not response 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. Accurate link failure detection might
   help ingress RBridges to 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 [MoFRR].

5.2.1. Failure Detection

Egress RBridges (merge points) SHOULD realize the link failure as
early as possible so that failure affected egress RBridges may update
their RPF filters quickly to minimize the traffic disruption. Three
options are provided as follows.

1. Egress RBridges assume a minimum known packet rate for a given
   data stream [MoFRR]. A failure detection timer Td are 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.

2. With multi-destination BFD, when a link failure happens, affected
   egress RBridges can detect a lack of connectivity from the ingress
   [mBFD]. 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 these multicast packets. This is achieved by
the activation of only a single RPF filter. In normal case, egress
RBridges activate the primary RPF filter. When a link on the pruned
primary DT fails, ingress RBridge cannot reach some of the receivers.
When these unreachable receivers realize it, they SHOULD update their
RPF filters to receive packets sent on the backup DT.

5.3. Local Protection

In the local protection, the Point of Local Repair (PLR) happens at
the upstream RBridge connecting the failed link. It is this RBridge

that makes the decision to replicate the multicast traffic to recover
this link failure. Local protection can further save the time spent
on failure notification through the flooding of LSPs across the
campus. In addition, the failure detection can be speeded up using
[rbBFD], therefore local protection can minimize the service
disruption within 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 descendants of the ingress point on the distribution tree.
Moreover, distribution trees in TRILL are bidirectional and do not
share the same root. There are fundamental differences between the
distribution tree calculation of TRILL and those used in PIM and
mLDP, therefore local protection mechanisms used for PIM and mLDP,
such as [mMRT] and [MoFRR], are not applicable here.

5.3.1. Start Using 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 of the multicast frame MUST
remain unchanged. This is a halfway change of the DT for multicast
packets. Afterwards, the PLR begins to forward multicast traffic
along the backup DT. This is a change from [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, if PLR RB1 decides to send replicated multicast
packets according to the backup DT, it will send it 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 in the above local protection is put together to get
a whole "walk through" below.

In the normal case, multicast frames ingressed by RB7 with pruned
distribution on 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. When 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 on to the local link.

5.4. Switching Back to the Primary Distribution Tree

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

   The propagation of LSPs around the campus takes time. For safety, we
   assume all RBridges in the campus have converged to the new primary
   DT at t1+Ts. By default, Ts (the "settling time") is set to 30s but
   is configurable. 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
   according to 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. It MUST wait for another Ts,
      during which time period all RBridges converge to the new backup
      DT. At t1+3*Ts, the ingress RBridge MAY 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. Security Considerations

This document raises no new security issues for TRILL.

For general TRILL Security Considerations, see [RFC6325].

7. IANA Considerations

No new registry or registry entries are requested to be assigned by IANA. The Affinity Sub-TLV has already been defined in [6326bis]. This document does not change its definition. RFC Editor: please remove this section before publication.

Acknowledgements

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

8. References

8.1. Normative References

[6326bis] D. Eastlake, T. Senevirathne, et al., "Transparent
          Interconnection of Lots of Links (TRILL) Use of IS-IS",
          draft-ietf-isis-rfc6326bis-01.txt, work in Progress.

[CMT]     T. Senevirathne, J. Pathangi, et al, "Coordinated Multicast
          Trees (CMT) for TRILL", draft-ietf-trill-cmt-02.txt, work
          in progress.

[RFC6325] R. Perlman, D. Eastlake, et al, "RBridges: Base Protocol
          Specification", RFC 6325, July 2011.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,
          "Protocol Independent Multicast - Sparse Mode (PIM-SM):
          Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas,
          "Label Distribution Protocol Extensions for Point-to-
          Multipoint and Multipoint-to-Multipoint Label Switched
          Paths", RFC 6388, November 2011.

[rbBFD]   V. Manral, D. Eastlake, et al, "TRILL (Transparent
          Interconnetion of Lots of Links): Bidirectional Forwarding
          Detection (BFD) Support", draft-ietf-trill-rbridge-bfd-
          07.txt, work in progress.

[ClearC]  Eastlake, D., M. Zhang, A. Ghanwani, V. Manral, A.

Banerjee, "TRILL: Clarifications, Corrections, and Updates"
draft-ietf-trill-clear-correct, in RFC Editor's queue.

8.2. Informative References

[mMRT]      A. Atlas, R. Kebler, et al., "An Architecture for Multicast
Protection Using Maximally Redundant Trees", draft-atlas-
rtgwg-mrt-mc-arch-02.txt, work in progress.

[MoFRR]     A. Karan, C. Filsfils, et al., "Multicast only Fast Re-
Route", draft-ietf-rtgwg-mofrr-02.txt, work in progress.

[mBFD]      D. Katz, D. Ward, "BFD for Multipoint Networks", draft-
ietf-bfd-multipoint-02.txt, work in progress.

[FGL]       D. Eastlake, M. Zhang, P. Agarwal, R. Perlman, D. Dutt,
"TRILL (Transparent Interconnection of Lots of Links):
Fine-Grained Labeling", draft-ietf-trill-fine-labeling, in
RFC Editor's queue.

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
   Cisco Systems
   375 East Tasman Drive,
   San Jose, CA 95134

   Phone: +1-408-853-2291
   Email: tsenevir@cisco.com

   Janardhanan Pathangi
   Dell/Force10 Networks
   Olympia Technology Park,
   Guindy Chennai 600 032

   Phone: +91 44 4220 8400
   Email: Pathangi_Janardhanan@Dell.com

   Ayan Banerjee
   Insieme Networks
   210 W Tasman Dr,
   San Jose, CA 95134

   Email: ayabaner@gmail.com

   Anoop Ghanwani
   Dell
   350 Holger Way
   San Jose, CA 95134

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

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

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