TRILL Working Group
INTERNET-DRAFT
Intended status: Proposed Standard
Updates: 6325, 6327, 6439

Donald Eastlake
Mingui Zhang
Huawei
Anoop Ghanwani
Dell
Ayan Banerjee
Cisco
Vishwas Manral
Hewlett-Packard

Expires: July 26, 2012

January 27, 2012

TRILL: Clarifications, Corrections, and Updates
<draft-ietf-trill-clear-correct-00.txt>

Abstract

The IETF TRILL (TRansparent Interconnection of Lots of Links) protocol provides least cost pair-wise data forwarding without configuration in multi-hop networks with arbitrary topology, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS (Intermediate System to Intermediate System) link state routing and by encapsulating traffic using a header that includes a hop count. Since the TRILL base protocol was approved in March 2010, active development of TRILL has revealed a few errata in the original RFC 6325 and some cases that could use clarifications or updates.

RFC 6327, RFC 6439, and RFC XXXX, provide clarifications with respect to Adjacency, Appointed Forwarders, and the TRILL ESADI protocol. This document provide other known clarifications, corrections, and updates to RFC 6325, RFC 6327, and RFC 6439.

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

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] provides optimal pair-wise data frame forwarding without configuration in multi-hop networks with arbitrary topology, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS (Intermediate System to Intermediate System) [IS-IS] [RFC1195] [RFC6326bis] link state routing and encapsulating traffic using a header that includes a hop count. The design supports VLANs (Virtual Local Area Networks) and optimization of the distribution of multi-destination frames based on VLANs and IP derived multicast groups.

Since the TRILL base protocol [RFC6325] was approved, the active development of TRILL has revealed a few errors in the original specification document [RFC6325] and cases that could use clarifications or updates.

[RFC6327], [RFC6439], and [RFCXXXX], provide clarifications with respect to Adjacency, Appointed Forwarders, and the TRILL ESADI protocol. This document provides other known clarifications, corrections, and updates to [RFC6325], [RFC6327], and [RFC6439].

1.1 Precedence

In case of conflict between this document and any of [RFC6325], [RFC6327], or [RFC6439], this document takes precedence. In addition, Section 1.2 (Normative Content and Precedence) of [RFC6325] is updated to provide a more complete precedence ordering of the sections of [RFC6325] as following, where sections to the left take precedence over sections to their right:

4 > 3 > 7 > 5 > 2 > 6 > 1

1.2 Terminology and Acronyms

This document uses the acronyms defined in [RFC6325] and the following additional acronyms:

CFI - Canonical Format Indicator [802]

DEI - Drop Eligibility Indicator [802.1Q-2011]

OOMF - Overload Originated Multi-destination Frame

TRILL Switch - An alternative name for an RBridge

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

2. Overloaded and/or Unreachable RBridges

RBridges may be in overload as indicated by the [IS-IS] overload flag in their LSPs. This means that either (1) they are incapable of holding the entire link state database and thus do not have a view of the entire topology or (2) they have been configured to have the overload bit on. Although networks should be engineered to avoid actual link state overload, it might occur under various circumstances. For example, if a large campus included one or more low-end TRILL Switches.

It is a common operational practice to set the overload bit in an [IS-IS] router (such as an RBridge) when performing maintenance on that router that might affect its ability to correctly forward frames; this will usually leave the router reachable for maintenance traffic but transit traffic will not normally be routed through it. (Also, in some cases, TRILL provides for setting the overload bit in the pseudo node of a link to stop TRILL Data traffic on an access link (see Section 4.9.1 of [RFC6325]).)

[IS-IS] and TRILL make a reasonable effort to do what they can even if some RBridges/routers are in overload. They can do reasonable well if a few scattered nodes are in overload. However, actual least cost paths are no longer assured if any RBridges are in overload.

For the effect of overload on the appointment of forwarders, see Section 10.2.

In this Section 2, the term "neighbor" refers only to actual RBridges and ignores psuedo nodes.

2.1 Reachability

Frames are not least cost routed through an overloaded TRILL Switch if any other path is available, although they may originate or terminate at an overloaded TRILL Switch. In addition, frames will not be least cost routed over links with cost 2**24 - 1; such links are reserved for traffic engineered frames the handling of which is beyond the scope of this document.

As a result, a portion of the campus may be unreachable for least cost routed TRILL Data because all paths to it would be through a link with cost 2**24 - 1. For example, an RBridge RB1 is not reachable by TRILL Data if all of its neighbors are connected to RB1 by links with cost 2**24 - 1. Such RBridges are called "data unreachable".

The link state database at an RBridge RB1 can also contain

information on TRILL Switches that are unreachable by IS-IS link state flooding due to link or RBridge failures. When such failures partition the campus, the TRILL Switches adjacent to the failure and on the same side of the failure as RB1 will update their LSPs to show the lack of connectivity and RB1 will receive those updates. However, LSPs held by RB1 for TRILL Switches on the far side of the failure will not be updated and may stay around until they time out, which could be tens of seconds or longer. As a result, RB1 will be aware of the partition. Nodes on the far side of the partition are both "IS-IS unreachable" and data unreachable.

2.2 Distribution Trees

A RBridge in overload cannot be trusted to correctly calculate distribution trees or correctly perform the Reverse Path Forwarding Check. Therefore, it cannot be trusted to forward multi-destination TRILL Data frames. It can only appear as a leaf node in a TRILL multi-destination distribution tree. Furthermore, if all the immediate neighbors of an RBridge are overloaded, then it is omitted from all trees in the campus and is unreachable by multi-destination frames.

When an RBridge determines what nicknames to use as the roots of the distribution trees it calculates, it MUST ignore all nicknames held by TRILL Switches that are in overload or are data unreachable. When calculating Reverse Path Forwarding Checks for multi-destination frames, an RBridge RB1 can similarly ignore any trees that cannot reach to RB1 even if other RBridges list those trees as trees those other TRILL Switches might use. (But see Section 3.)

2.3 Overloaded Receipt of TRILL Data Frames

The receipt of TRILL Data frames by overloaded RBridge RB2 is discussed in the subsections below. In all cases, the normal Hop Count decrement is performed and the TRILL Data frame is discarded if the result is less than one or if the egress nickname is illegal.

2.3.1 Known Unicast Receipt

RB2 will not usually receive unicast TRILL Data frames unless it is the egress, in which case it decapsulates and delivers the frames normally. If RB2 receives a unicast TRILL Data frame for which it is not the egress, perhaps because a neighbor does not yet know it is in

overload, RB2 MUST NOT discard the frame because the egress is an

unknown nickname as it might not know about all nicknames due to its overloaded condition. If any neighbor, other than the neighbor from which it received the frame, is not overloaded it MUST attempt to forward the frame to one of those neighbors. If there is no such neighbor, the frame is discarded.

2.3.2 Multi-Destination Receipt

If RB2 in overload receives a multi-destination TRILL Data frame, RB2 MUST NOT apply a Reverse Path Forwarding Check since, due to overload, it might not do so correctly. RB2 decapsulates and delivers the frame locally where it is Appointed Forwarder for the frame's VLAN, subject to any multicast pruning. But since, as stated above, RB2 can only be the leaf of a distribution tree, it MUST NOT forward a multi-destination TRILL Data frame (except as an egressed native frame where RB2 is Appointed Forwarder).

2.4 Overloaded Origination of TRILL Data Frames

Overloaded origination of unicast frames with known egress and of multi-destination frames are discussed in the subsections below.

2.4.1 Known Unicast Origination

When an overloaded RBridge RB2 ingresses or creates a known destination unicast TRILL Data frame, it delivers it locally if the destination MAC is local. Otherwise RB2 unicasts it to any neighbor TRILL Switch that is not overloaded. It MAY use what routing information it has to help select the neighbor.

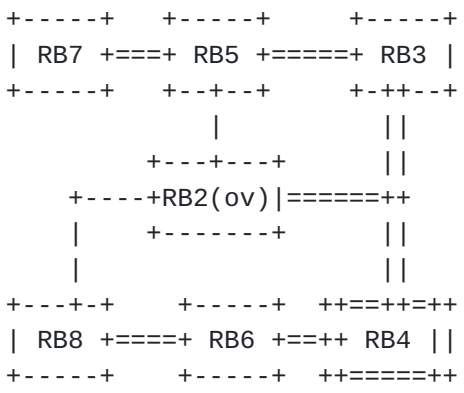
2.4.2 Multi-Destination Origination

Overloaded RBridge RB2 ingressing or creating a multi-destination TRILL Data frame is more complex than for a known unicast frame.

2.4.2.1 An Example Network

For example, consider the network below in which, for simplicity, end stations and any bridges are not shown. There is one distribution tree of which RB4 is the root and which is represented by double

lines. Only RBridge RB2 is overloaded.



Since RB2 is overloaded it does not know what the distribution tree or trees are for the network. Thus there is no way it can provide normal TRILL Data encapsulation for multi-destination native frames. So RB2 tunnels the frame to a neighbor that is not overloaded if it has such a neighbor that signal it is willing to offer this service. RBridges indicate this in their Hellos as described below. This service is called OOMF (Overloaded Origination of Multi-destination Frame) service.

- The multi-destination frame MUST NOT be locally distributed in native form at RB2 before tunneling to a neighbor because this would cause the frame to be delivered twice. For example, if RB2 locally distributed a multicast native frame and then tunneled it to RB5, RB2 would get a copy of the frame when RB3 transmitted it as a TRILL Data frame on the multi-access RB2-RB3-RB4 link. Since RB2 would, in general, not be able to tell that this was a frame it had tunneled for distribution, RB2 would decapsulate it and locally distribute it a second time.

- On the other hand, if there is no neighbor of RB2 offering RB2 the OOMF service, RB2 cannot tunnel the frame to a neighbor. In this case RB2 MUST locally distribute the frame where it is Appointed Forwarder for the frame's VLAN and optionally subject to multicast pruning.

2.4.2.2 Indicating OOMF Support

A RBridge RB3 indicates its willingness to offer the OOMF service to RB2 in the TRILL Neighbor TLV in RB3's TRILL Hellos by setting a bit associated with the SNPA (MAC address) of RB2 on the link. (See Section 11.) Overloaded RBridge RB2 can only distribute multi-destination TRILL Data frames to the campus if a neighbor of RB2 not in overload offers RB2 the OOMF service. If RB2 does not have OOMF

service available to it, RB2 can still receive multi-destination

frames from non-overloaded neighbors and, if RB2 should originate or ingress such a frame, it distributes it locally in native form.

2.4.2.3 Using OOMF Service

If RB2 sees this OOMF (Overloaded Origination of Multi-destination Frame) service advertised for it by any of its neighbors on any link to which RB2 connects, it selects one such neighbor by a means beyond the scope of this document. Assuming RB2 selects RB3 to handle multi-destination frames it originates. RB2 MUST advertise in its LSP that it might use any of the distribution trees that RB3 advertises it might use so that the Reverse Path Forwarding Check will work in the rest of the campus. Thus, notwithstanding its overloaded state, RB2 MUST retain this information from RB3 LSPs, which it will receive as it is directly connected to RB3.

RB2 then encapsulates such frames as TRILL Data frames to RB3 as follows: M bit = 0, Hop Count = 2, ingress nickname = a nickname held by RB2, and, since RB2 cannot tell what distribution tree RB3 will use, egress nickname = a special nickname indicating an OOMF frame (see Section 11). RB2 then unicasts this TRILL Data frame to RB3. (Implementation of Item 4 in Section 4 below provides reasonable assurance that, notwithstanding its overloaded state, the ingress nickname used by RB2 will be unique within at least the portion of the campus that is IS-IS reachable from RB2.)

On receipt of such a frame, RB3 does the following:

- change the egress nickname field to designate a distribution tree that RB3 normally uses,
- set the M bit to one,
- change the Hop Count to the value it would normally use if it were the ingress, and
- forward the frame on that tree.

RB3 MAY rate limit the number of frames for which it is providing this service by discarding some such frames from RB2. The provision of even limited bandwidth for OOMFs by RB3, perhaps via the slow path, may be important to the bootstrapping of services at RB2 or at end stations connected to RB, such as supporting DHCP and ARP/ND. (Everyone sometimes needs a little OOMF (pronounced oompf) to get off the ground.)

3. Distribution Trees

A correction and a clarification related to distribution trees appear in the subsections below. See also Section 2.2.

3.1 Number of Distribution Trees

In [RFC6325], Section 4.5.2, page 56, Point 2, 4th paragraph, the parenthetical "(up to the maximum of {j,k})" is incorrect. It should read "(up to k if j is zero or the minimum of (j, k) if j is non-zero)".

3.2 Distribution Tree Updates

When a link state database change causes a change in the distribution tree(s), there are several possibilities. If a tree root remains a tree root but the tree changes, then local forwarding and RPFC entries for that tree should be updated as soon as practical. Similarly, if a new nickname becomes a tree root, forwarding and RPFC entries for the new tree should be installed as soon as practical. However, if a nickname ceases to be a tree root and there is sufficient room in local tables, the forwarding and RPFC entries for the former tree MAY be retained so that any multi-destination TRILL Data frames already in flight on that tree have a higher probability of being delivered.

4. Nickname Selection

Nickname selection is covered by Section 3.7.3 of [RFC6325]. However, the following should be noted:

1. The second sentence in the second bullet item in Section 3.7.3 of [RFC6325] on page 25 is erroneous and is corrected as follows:

   1.a The occurrence of "IS-IS ID (LAN ID)" is replaced with "priority".

   1.b The occurrence of "IS-IS System ID" is replaced with "seven byte IS-IS ID (LAN ID)".

   The resulting corrected [RFC6325] sentence reads as follows: "If RB1 chooses nickname x, and RB1 discovers, through receipt of an LSP for RB2 at any later time, that RB2 has also chosen x, then the RBridge or pseudonode with the numerically higher priority keeps the nickname, or if there is a tie in priority, the RBridge with the numerically higher seven byte IS-IS ID (LAN ID) keeps the nickname, and the other RBridge MUST select a new nickname."

2. In examining the link state database for nickname conflicts, nicknames held by IS-IS unreachable TRILL Switches MUST be ignored but nicknames held by IS-IS reachable TRILL Switches MUST NOT be ignored even if they are data unreachable.

3. An RBridge may need to select a new nickname, either initially because it has none or because of a conflict. When doing so, the RBridge MUST consider as available all nicknames that do not appear in its link state database or that appear to be held by IS-IS unreachable TRILL Switches; however, it SHOULD give preference to selecting new nicknames that do not appear to be held by any TRILL Switch in the campus, reachable or unreachable, so as to minimize conflicts if IS-IS unreachable TRILL Switches later become reachable.

4. An RBridge, even after it has acquired a nickname for which there appears to be no conflicting claimant, MUST continue to monitor for conflicts with the nickname or nicknames it holds. It does so by checking in LSPs it receives that should update its link state database for any of its nicknames held with higher priority by another TRILL Switch that is IS-IS reachable. If it finds such a conflict, it MUST select a new nickname. (It is possible to receive an LSP that should update the link state database but does not due to overload.)

5. In the very unlikely case that an RBridge is unable to obtain a nickname because all valid nicknames (0x0001 through 0xFFBF

inclusive) are in use with higher priority by IS-IS reachable

TRILL Switches, it will be unable to act as an ingress, egress, or tree root but will still be able to function as a transit TRILL Switch. Although it cannot be a tree root, such an RBridge is included in distribution trees computed for the campus unless all its neighbors are overloaded. It would not be possible to send an RBridge Channel message to such a TRILL Switch [Channel].

5. MTU (Maximum Transmission Unit)

MTU values in TRILL key off the originatingL1LSPBufferSize value communicated in the IS-IS originatingLSPBufferSize TLV [IS-IS]. The campus-wide value Sz, as described in [RFC6325] Section 4.3.1, is the minimum value of originatingL1LSPBufferSize for the RBridges in a campus, but not less than 1470. The MTU testing mechanism and limiting LSPs to Sz assures that the LSPs can be flooded properly by IS-IS and thus that IS-IS can operate properly.

If nothing is known about the campus, the originatingL1LSPBufferSize for an RBridge should default to the minimum of the LSP size that its TRILL IS-IS software can handle and the minimum MTU of the ports that it might use to receive or transmit LSPs. However, to avoid having to refragment LSPs, originatingL1LSPBufferSize SHOULD be configured to a smaller value if it is known that other RBridges will be announcing such smaller value or that the campus will partition due to a significant number of links with an MTU of such smaller value. In a well configured campus, to minimize any LSP re-sizing, it is desirable for all RBridges to be configured with the same originatingL1LSPBufferSize.

Section 5.1 below corrects errata in [RFC6325] and Section 5.2 clarifies the meaning of various MTU (Maximum Transmission Unit) limits for TRILL Ethernet links.

5.1 MTU Related Errata in RFC 6325

Three MTU related errata in [RFC6325] are corrected in the subsections below.

5.1.1 MTU PDU Addressing

Section 4.3.2 of [RFC6325] incorrectly states that multi-destination MTU-probe and MTU-ack TRILL IS-IS PDUs are sent on Ethernet links with the All-RBridges multicast address as the Outer.MacDA. As TRILL IS-IS PDUs, when multicast on an Ethernet link, they MUST be sent to the All-IS-IS-RBridges multicast address.

5.1.2 MTU PDU Processing

As discussed in [RFC6325] and, in more detail, in [RFC6327], MTU-probe and MTU-ack PDUs MAY be unicast; however, Section 4.6 of

[RFC6325] erroneously does not allow for this possibility. It is

corrected by replacing Item numbered "1" in Section 4.6.2 of [RFC6325] with the following quoted text to which TRILL Switches MUST conform:

"1. If the Ethertype is L2-IS-IS and the Outer.MacDA is either All-IS-IS-RBridges or the unicast MAC address of the receiving RBridge port, the frame is handled as described in Section 4.6.2.1"

The reference to "Section 4.6.2.1" in the above quoted text is to that Section in [RFC6325].

5.1.3 MTU Testing

The last two sentences of Section 4.3.2 of [RFC6325] have errors. They currently read:

> If X is not greater than Sz, then RB1 sets the "failed minimum MTU test" flag for RB2 in RB1's Hello. If size X succeeds, and X > Sz, then RB1 advertises the largest tested X for each adjacency in the TRILL Hellos RB1 sends on that link, and RB1 MAY advertise X as an attribute of the link to RB2 in RB1's LSP.

They should read:

> If X is not greater than or equal to Sz, then RB1 sets the "failed minimum MTU test" flag for RB2 in RB1's Hello. If size X succeeds, and X >= Sz, then RB1 advertises the largest tested X for each adjacency in the TRILL Hellos RB1 sends on that link, and RB1 MAY advertise X as an attribute of the link to RB2 in RB1's LSP.

5.2 Ethernet MTU Values

originatingL1LSPBufferSize is the maximum permitted size of LSPs after the eight byte fixed IS-IS PDU header. This IS-IS PDU header starts with the 0x83 Intradomain Routeing Protocol Discriminator byte and ends with the Maximum Area Addresses byte, inclusive. In layer 3 IS-IS, originatingL1LSPBufferSize defaults to 1492 bytes and thus the default Layer 3 LSP size, including this header, is 1500 bytes. In TRILL, originatingL1LSPBufferSize defaults to 1470 bytes, allowing 22 bytes of additional headroom or safety margin to accommodate legacy devices with, for example, the classic Ethernet maximum MTU, and headers such as an Outer.VLAN. We will call this safety margin "Margin" below.

Assuming the campus wide minimum link MTU is Sz, RBridges on Ethernet

links MUST limit most TRILL IS-IS PDUs so that PDUz (the length of the PDU starting just before and including the L2-IS-IS Ethertype and ending just before the Ethernet frame FCS) does not to exceed

PDUz = (Sz + 32 - Margin) bytes

The PDU exceptions are TRILL Hello PDUs, which MUST NOT exceed this limit assuming an Sz of 1470 bytes, and MTU-probe and MTU-ack PDUs which are padded, depending on the size Tz being tested, to (Tz + 32 - Margin) bytes.

Sz does not limit TRILL Data frames. They are only limited by the MTU of the RBridges and links that they actually pass through; however, links that can accommodate IS-IS PDUs up to Sz should accommodate, with a reasonable safety margin, TRILL Data frame payloads, starting after the Inner.VLAN and ending just before the FCS, of (Sz + 10 - Margin) bytes. Most modern Ethernet equipment has ample headroom for frames with extensive headers and is sometimes engineered to accommodate 9K byte jumbo frames.

6. Port Modes

Section 4.9.1 of [RFC6325] specifies four mode bits for RBridge ports but may not be completely clear on the effects of various combinations of bits.

The table below explicitly indicates the effect of all possible combinations of the TRILL port mode bits. "*" in one of the first four columns indicates that the bit can be either zero or one. The following columns indicate allowed frame types. The Disable bit normally disables all frames but, as an implementation choice, some or all low level Layer 2 control frames (a specified in [RFC6325] Section 1.4) can still be sent or received.

Disable	Active	Transparent	Learning	Layer 2 Control	native ingress egress	LSP SNP MTU	TRILL Hello	P2P Hello
0	0	0	0	Yes	Yes	Yes	Yes	No
0	0	0	1	Yes	No	Yes	Yes	No
0	0	1	0	Yes	Yes	No	Yes	No
0	0	1	1	Yes	No	No	Yes	No
0	1	0	*	Yes	No	Yes	No	Yes
0	1	1	*	Yes	No	No	No	Yes
1	*	*	*	Optional	No	No	No	No

7. The CFI / DEI Bit

In May 2011, the IEEE promulgated [802.1Q-2011] which changes the meaning of the bit between the priority and VLAN ID bits in the payload of C-VLAN tags. Previously this bit was called the CFI (Canonical Format Indicator) bit [802] and had a special meaning in connection with IEEE 802.5 (Token Ring) frames. Now, under [802.1Q-2011], it is a DEI (Drop Eligibility Indicator) bit, similar to that bit in S-VLAN / B-VLAN tags where this bit has always been a DEI bit.

The TRILL base protocol specification [RFC6325] assumed, in effect, that the link by which end stations are connected to TRILL Switches and the virtual link provided by the TRILL Data frame are IEEE 802.3 Ethernet links on which the CFI bit is always zero. Should an end station be attached by some other type of link, such as a Token Ring link, [RFC6325] implicitly assumed that such frames would be canonicalized to 802.3 frames before being ingressed and similarly, on egress, such frames would be converted from 802.3 to the appropriate frame type for the link. Thus, [RFC6325] required that the CFI bit in the Inner.VLAN always be zero.

However, for TRILL Switches with ports conforming to the change incorporated in the IEEE 802.1Q-2011 standard, the bit in the Inner.VLAN, now a DEI bit, MUST be set to the DEI value provided by the EISS interface on ingressing a native frame. Similarly, this bit MUST be provided to the EISS when transiting or egressing a TRILL Data frame. As with the 3-bit priority field, the DEI bit to use in forwarding a transit frame MUST be taken from the Inner.VLAN. The exact effect on the Outer.VLAN DEI and priority bits and whether or not an Outer.VLAN appears at all on the wire for output frames may depend on output port configuration.

TRILL Switch campuses with a mixture of ports, some compliant with [802.1Q-2011] and some compliant with pre-802.1Q-2011 standards, especially if they have actual Token Ring links, may operate incorrectly and may corrupt data, just as a bridged LAN with such mixed bridges and ports would.

8. Graceful Restart

TRILL Switches SHOULD support the features specified in [RFC5306] which describes a mechanism for a restarting IS-IS router to signal to its neighbors that it is restarting, allowing them to reestablish their adjacencies without cycling through the down state, while still correctly initiating link state database synchronization.

9. Some Updates to RFC 6327

[RFC6327] provides for multiple states of the potential adjacency between two TRILL Switches. It makes clear that only an adjacency in the "Report" state is reported in LSPs. LSP synchronization (LSP and SNP transmission and receipt), however, is performed if and only if there is at least one adjacency on the link in the "Two-Way" or "Report" state.

To support the PORT-TRILL-VER sub-TLV specified in [RFC6326bis], the following updates are made to [RFC6327]:

1. The paragraph immediately before the 3.2 section header is modified by adding "TRILL-PORT-VER sub-TLV [RFC6326bis] if included" to those items which MUST be the same in all TRILL Hellos sent out the same RBridge port regardless of the VLAN on which they are sent but can occasionally change.

2. In Section 3.2, the state entry for each adjacency is expanded to include the 5 bytes of data from the TRILL-PORT-VER received in the most recent TRILL Hello from the remote RBridge.

3. In Section 3.3, a bullet item as follows is added to the bullet items after the event descriptions: "The five bytes of TRILL-PORT-VER data are set from that sub-TLV in the Hello or set to zero if that sub-TLV does not occur in the Hello."

4. In the first part of Section 4, a bullet item is added to the list as follows: "The five bytes of TRILL-PORT-VER sub-TLV data used in TRILL Hellos sent on the port."

10. Updates on Appointed Forwarders and Inhibition

An optional method of Hello reduction is specified in Section 10.1 below and a recommendation on forwarder appointments in the face of overload is given in Section 10.2.

10.1 Optional TRILL Hello Reduction

If a network manager has sufficient confidence that they know the configuration of bridges, ports, and the like, within a link, they may be able to reduce the number of TRILL Hellos sent on that link; for example, if all RBridges on the link will see all Hellos regardless of VLAN constraints, Hellos could be sent on fewer VLANs. However, because adjacencies are established in the Designated VLAN, an RBridge MUST always attempt to send Hellos in the Designated VLAN. Hello reduction makes TRILL less robust in the face of partitioned VLANs or disagreement over the Designated VLAN or the like in a link; however, as long as all RBridge ports on the link are configured for the same desired Designated VLAN, can see each others frames in that VLAN, and utilize the mechanisms specified below to update VLAN inhibition timers, operations will be safe. (These considerations do not arise on links between RBridges that are configured as point-to-point since, in that case, each RBridge sends point-to-point Hellos, other TRILL IS-IS PDUs, and TRILL Data frames only in what it believes to be the Designated VLAN of the link and no native frame end station service is provided.)

The provision for a configurable set of "Announcing VLANs", as described in Section 4.6.3 of [RFC6325] provides a mechanism in the TRILL base protocol for a reduction in TRILL Hellos.

To maintain loop safety in the face of occasional lost frames, RBridge failures, link failures, new RBridges coming up on a link, and the like, the inhibition mechanism specified in [RFC6439] is still required. Under Section 3 of [RFC6439], a VLAN inhibition timer can only be set by the receipt of a Hello sent or received in that VLAN. Thus, to safely send a reduced number of TRILL Hellos on a reduced number of VLANs requires additional mechanisms to set the VLAN inhibition timers at an RBridge, thus extending Section 3, Item 4, of [RFC6439]. Two such mechanisms are specified below. Support for both of these mechanisms is indicated by a capability bit in the TRILL-PORT-VER sub-TLV (see Section 9 above and [RFC6326bis]). Unless all adjacencies that are not in the Down state out a port indicate support of these mechanisms and the mechanisms are used, it may be unsafe to reduce the VLANs on which TRILL Hellos are sent to fewer VLANs than recommended in [RFC6325].

1. An RBridge RB2 MAY include in any TRILL Hello an Appointed Forwarders sub-TLV [RFC6326bis] appointing itself for one or more ranges of VLANs. The Appointee Nickname field(s) in the Appointed Forwarder sub-TLV MUST be the same as the Sender Nickname in the Special VLANs and Flags sub-TLV in the TRILL Hello. This indicates the sending RBridge believes it is Appointed Forwarder for those VLANs. An RBridge receiving such a sub-TLV sets each of its VLAN inhibition timers for every VLAN in the block or blocks listed in the Appointed Forwarders sub-TLV to the maximum of its current value and the Holding Time of the Hello containing the sub-TLV. This is backwards compatible because such sub-TLVs will have no effect on any receiving RBridge not implementing this mechanism unless RB2 is the DRB sending Hello on the Designated VLAN in which case, as specified in [RFC6439], RB2 MUST include in the Hello all forwarder appointments, if any, for RBridges other than itself on the link.

2. An RBridge MAY use the new VLANs Appointed sub-TLV [RFC6326bis]. When RB1 receives a VLANs Appointed sub-TLV in a TRILL Hello from RB2 on any VLAN, RB1 updates the VLAN inhibition timers for all the VLANs that RB2 lists in that sub-TLV as VLANs for which RB2 is Appointed Forwarder. Each such timer is updated to the maximum of its current value and the Holding Time of the TRILL Hello containing the VLANs Appointed sub-TLV. This sub-TLV will be an unknown sub-TLV to RBridge not implementing it and such RBridges will ignore it. Even if a TRILL Hello send by the DRB on the Designated VLAN includes one or more VLANs Appointed sub-TLVs, as long as no Appointed Forwarders sub-TLVs appear, the Hello is not required to indicate all forwarder appointments.

Two different encoding are providing above to optimize the listing of VLANs. Large blocks of contiguous VLANs are more efficiently encoded with the Appointed Forwarders sub-TLV and scattered VLANs are more efficiently encoded with the VLANs Appointed sub-TLV. These encoding may be mixed in the same Hello and the use of these sub-TLVs does not affect the requirement that the "AF" bit in the Special VLANs and Flags sub-TLV MUST be set if the originating RBridge believes it is Appointed Forwarder for the VLAN in which the Hello is sent. If the above mechanisms are used on a link, then each RBridge on the link MUST send Hellos in one or more VLANs with such VLANs Appointed sub-TLV(s) and/or self-appointment Appointed Forwarders sub-TLV(s) and the "AF" bit appropriately set such that no VLAN inhibition timer will improperly expire unless three or more Hellos are lost. For example, an RBridge could announce all VLANs for which it believes it is Appointed Forwarder in a Hello sent on the Designated VLAN three times per Holding Time.

10.2 Overload and Appointed Forwarders

An RBridge in overload (see Section 2) will, in general, do a poorer job of ingressing and forwarding frames than an RBridge not in overload that has full knowldge of the campus topology. For example, an overloaded RBridge may not be able to distribute multi-destination TRILL Data frames at all.

Therefore, the DRB SHOULD NOT appointed an RBridge in overload as Appointed Forwarder for an VLAN unless there is no alternative. Furthermore, if an Appointed Forwarder becomes overloaded, the DRB SHOULD re-assign VLANs from the overloaded RBridged to another RBridge on the link that is not overloaded, if one is available.

A counter-example would be if all campus end stations in VLAN-x were on links attached to RB1 via ports where VLAN-x was enabled. In such a case, RB1 SHOULD be made the VLAN-x Appointed Forwarder on all such link even if RB1 is overloaded.

11. IANA Considerations

The following IANA actions are required:

1. The previously reserved nickname 0xTBD [0xFFC1 suggested] is allocated for use in the TRILL Header egress nickname field to indicate an Overload Originated Multi-destination Frame (OOMF).

2. Bit 1 from the seven previously reserved (RESV) bits in the per neighbor "Neighbor RECORD" in the TRILL Neighbor TLV [RFC6326bis] is allocated to indicate that the RBridge sending the TRILL Hello volunteers to provide the OOMF forwarding service described in Section 2.4.2 to such frames originated by the TRILL Switch whose SNPA (MAC address) appears in that Neighbor RECORD.

3. Bit 0 is allocated from the Capability bits in the TRILL-PORT-VER sub-TLV [RFC6326bis] to indicate support of the VLANs Appointed sub-TLV [RFC6326bis] and the VLAN inhibition setting mechanisms specified in Section 10.1.

12. Security Considerations

This memo improves the documentation of the TRILL protocol, corrects some errors in [RFC6325], and updates [RFC6325], [RFC6327], and [RFC6439]. It does not change the security considerations of these RFCs.

Acknowledgements

The contributions of the following persons are gratefully acknowledged:

Somnath Chatterjee, Weiguo Hao, Rakesh Kumar, Yizhou Li, Radia Perlman

This document was produced with raw nroff. All macros used were defined in the source file.

Normative References

[802.1Q-2011] - IEEE 802.1, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, May 2011.

[IS-IS] - ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.

[RFC1195] - Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[RFC2119] - Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5306] - Shand, M. and L. Ginsberg, "Restart Signaling for IS-IS", RFC 5306, October 2008.

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

[RFC6327] - Eastlake 3rd, D., Perlman, R., Ghanwani, A., Dutt, D., and V. Manral, "Routing Bridges (RBridges): Adjacency", RFC 6327, July 2011.

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

[RFC6326bis] - Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, draft-eastlake-isis-rfc6326bis, work in progress.

Informative References

[802] - IEEE 802, "IEEE Standard for Local and metropolitan area networks: Overview and Architecture", IEEE Std 802.1-2001, 8 March 2002.

[Channel] - draft-ietf-trill-rbridge-channel, work in progress.

[RFCXXXX] - H. Zhai, F. Hu, R. Perlman, D. Eastlake, "RBridges: The ESADI Protocol", draft-hu-trill-rbridge-esadi, work in progress.

Authors' Addresses

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

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

Mingui Zhang
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Anoop Ghanwani
Dell
350 Holger Way
San Jose, CA 95134 USA

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

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

Tel.: +1-408-527-0539
Email: ayabaner@cisco.com

Vishwas Manral
HP Networking
19111 Pruneridge Avenue
Cupertino, CA 95014 USA

Tel: +1-408-477-0000
Email: vishwas.manral@hp.com

Copyright and IPR Provisions