INTERNET-DRAFT
Intended Status: Standards Track
Updates: 7177

Mingui Zhang
Xudong Zhang
Donald Eastlake
Huawei
Radia Perlman
Intel
Vishwas Manral
Ionos
Somnath Chatterjee
Cisco

Expires: August 15, 2015
February 11, 2015

TRILL IS-IS MTU Negotiation
draft-zhang-trill-mtu-negotiation-08.txt

Abstract

The base IETF TRILL protocol has a TRILL campus wide MTU feature, specified in RFC 6325 and RFC 7177, that assures that link status changes can be successfully flooded throughout the campus while being able to take advantage of a campus wide capability to support jumbo packets. This document specifies optional updates to that MTU feature to take advantage, for appropriate link local packets, of link local MTUs that exceed the TRILL campus MTU. In addition, it specifies an efficient algorithm for local MTU testing. It updates RFC 7177.

Table of Contents

1. Introduction

[RFC6325] describes the way RBridges agree on the campus-wide minimum acceptable inter-RBridge MTU (Maximum Transmission Unit) size - the campus-wide "Sz" to ensure that link state flooding operates properly and all RBridges converge to the same link state. For the proper operation of TRILL IS-IS, all RBridges MUST format their LSPs not greater than the campus-wide Sz. [RFC7177] defines the diagram of state transitions of an adjacency. "Link MTU size is successfully tested" is part of an event (A6) causing the transition from "2-way" state to "Report" state for an adjacency. If MTU testing is enabled, this part means the link MTU testing of size X succeeds, and X is greater than or equal to the campus-wide Sz [RFC6325]. In other words, if this link cannot support an MTU of the campus-wide Sz, it will not be reported as part of the campus topology.

This document specifies a new value, link-wide "Lz" to represent the link-wide minimum acceptable inter-RBridge MTU size for a specific link. There are PDUs which are valid only to a local link, such as CSNPs and PSNPs. These PDUs should be formatted not greater than the link-wide Lz. Since link-wide Lz is frequently greater than the campus-wide Sz, link scope PDUs can, in such cases, be optionally formatted greater than the campus-wide Sz up to Lz.

An optional TRILL IS-IS MTU size testing algorithm is specified in Section 3 to detail the MTU testing method described in Section 4.3.2 of [RFC6325] and in [RFC7177]. Multicasting the MTU-probes is recommended when there are multiple RBridges on a link responding to the probing with MTU-ack [RFC7177]. The testing method and rules of this document are devised in a way to minimize the number of MTU probes for testing, which therefore reduces the number of multicast packets for MTU testing.

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].

2. Link-Wide TRILL IS-IS MTU Size

This document specifies a new value "Lz" to represent the acceptable inter-RBridge link MTU size on the local link. Link-wide Lz is the minimum Lz supported by all RBridges on a specific link. If the link is usable, Lz will be greater than or equal to the campus wide Sz MTU. Some TRILL IS-IS PDUs are exchanged only between neighbors instead of the whole campus. They should be confined by the link-wide Lz instead of the campus-wide Sz. CSNPs and PSNPs are examples of

such PDUs. They are exchanged just on the link as part of LSP synchronization.

[RFC7356] defines the PDUs which support flooding scopes in addition to area wide scope and domain wide scope. RBridges on a local link that support Lz greater than Sz MUST support the Extended L1 Circuit Scoped (E-L1CS) flooding. They use that flooding to exchange their maximally supportable value of "Lz". The smallest value of the Lz collected on a link, but not less than Sz, is the link-wide Lz. RBridge on a local link will be able to tell which other RBridges on that link support E-L1CS FS-LSPs because, as required by [RFC7180bis] all RBridges are required to include the Scoped Flooding Support TLV [RFC7356] in their TRILL Hellos.

The maximum sized level 1 link-local PDU, such as PSNP or CSNP, which may be generated by a system is controlled by the value of the management parameter originatingL1SNPBufferSize. This value determines Lz. The TRILL APPsub-TLV shown in Figure 2.1 SHOULD be included in a GENINFO TLV [RFC6823] in an E-L1CS-LSP number zero. If it is missing from a fragment zero E-L1CS-LSP or there is no fragment zero E-L1CS FS-LSP, it is assumed that its originating IS is implicitly advertising its originatingSNPBufferSize value as Sz octets.

E-L1CS FS-LSPs are link local and can also be sent up to Lz in size but, for robustness, E-L1CS fragment zero MUST NOT exceed 1470 bytes.

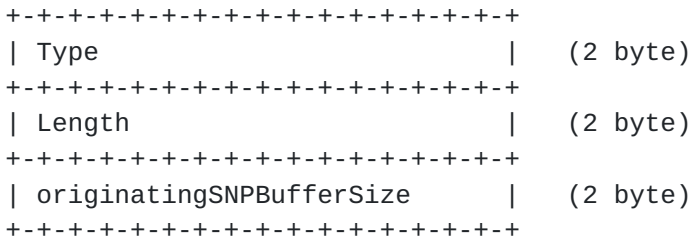

Figure 2.1: Lz is reported in the originatingSNPBufferSize TLV.

Type: set to originatingSNPBufferSize subTLV (TRILL APPsub-TLV type tbd). Two bytes because this APPsub-TLV appears in an Extended TLV [RFC7356].

Length: set to 2.

originatingSNPBufferSize: the local value of originatingL1SNPBufferSize, limited to 1470 to 65,535 bytes.

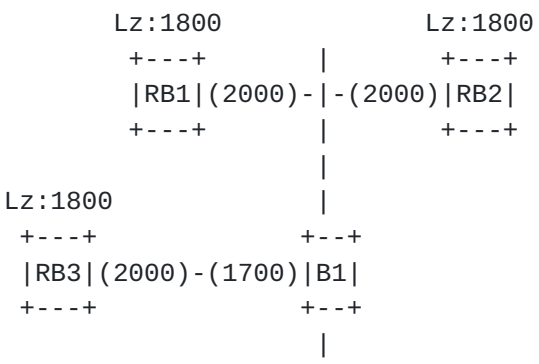

Figure 2.2: Link-wide Lz = 1800 v.s. tested link MTU size = 1700

Even if all RBridges on a specific link have reached consensus on the value of link-wide Lz, it does not mean that these RBridges can safely exchange PDUs between each other. Figure 2.2 shows such a corner case. RB1, RB2 and RB3 are three RBridges on the same link and their Lz is 1800, so the link-wide Lz of this link is 1800. There is an intermediate bridge (say B1) between RB2 and RB3 whose port MTU size is 1700. If RB2 sends PDUs formatted in chunk of size 1800, it will be discarded by B1.

Therefore the link MTU size should be tested. After the link MTU size of an adjacency is successfully tested, those link local PDUs such as CSNP, PSNP and also E-L1CS FS-LSP will be formatted no greater than the tested link MTU size and will be safely transmitted on this link.

As for campus-wide Sz, RBridges continue to propagate their originatingL1LSPBufferSize across the campus through the advertisement of LSPs as defined in Section 4.3.2 of [RFC6325]. The smallest value of Sz advertised by any RBridge, but not less than 1470, will be deemed as the campus-wide Sz. Each RBridge should format their "campus-wide" PDUs, for example LSPs, not greater than what they believe to be the campus-wide Sz.

3. Link MTU Size Testing

[RFC7177] defines the event A6 as including "MTU test is successful" if the MTU testing is enabled. As described in Section 4.3.2 of [RFC6325], this is a combination of the following event and condition.

Event: The link MTU size has been tested.

Condition: The link can support the campus-wide Sz.

This condition can be efficiently tested by the following "Binary Search Algorithm" and rules. The MTU-probe and MTU-ack PDUs are

specified in Section 3 of [RFC7176].

X, X1, and X2 are local integer variables.

Step 0: RB1 sends an MTU-probe padded to the size of link-wide Lz.

1) If RB1 successfully receives the MTU-ack from RB2 to the probe of the value of link-wide Lz within k tries (where k is a configurable parameter whose default is 3), then link MTU size is set to the size of link-wide Lz and stop.

2) RB1 tries to send an MTU-probe padded to the size 1470.

   a) If RB1 fails to receive an MTU-ack from RB2 after k tries, RB1 sets the "failed minimum MTU test" flag for RB2 in RB1's Hello and stop.

   b) Link MTU size <-- 1470, X1 <-- 1470, X2 <-- link-wide Lz, X <-- [(X1 + X2)/2] (Operation "[...]" returns the fraction-rounded-up integer.). Repeat Step 1.

Step 1: RB1 tries to send an MTU-probe padded to the size X.

1) If RB1 fails to receive an MTU-ack from RB2 after k tries, then:

   X2 <-- X and X <-- [(X1 + X2)/2]

2) If RB1 receives an MTU-ack to a probe of size X from RB2 then:

   link MTU size <-- X, X1 <-- X and X <-- [(X1 + X2)/2]

3) If X1 >= X2 or Step 1 has been repeated n times (where n is a configurable parameter whose default value is 5), stop. Else repeat Step 1.

MTU testing is only done in the Designated VLAN [RFC7177]. Since the execution of the above algorithm can be resource consuming, it is recommended that the DRB take the responsibility to do the testing. Multicast should be used instead of unicast when multiple RBridges are desired to respond with MTU-ack on the link. The Binary Search Algorithm is proposed here to minimize the probing attempts; therefore it reduces the number of multicast packets for MTU-probing.

The following rules are designed to determine whether the aforementioned "Condition" holds.

RBridges have figured out the upper bound (X2) and lower bound (X1) for the link MTU size from the execution of the above algorithm. If

the campus-wide Sz is smaller than the lower bound or greater than the upper bound, RBridges can directly judge whether the link supports the campus-wide Sz without MTU-probing.

(a) If X1 >= campus-wide Sz. This link can support campus-wide Sz.

(b) Else if X2 <= campus-wide Sz. This link cannot support campus-wide Sz.

Otherwise, RBridges need to test whether the link can support campus-wide Sz:

(c) X1 < campus-wide Sz < X2. RBridges need probe the link with MTU-probe messages padded to campus-wide Sz. If an MTU-ack is received within k tries, this link can support campus-wide Sz. Otherwise, this link cannot support campus-wide Sz. Through this test, the lower bound and upper bound of link MTU size can be updated accordingly.

4. Refreshing Campus-Wide Sz

RBridges may join or leave the campus, which may change the campus-wide Sz. The following recommendations are specified for refreshing the campus-wide Sz.

1) When a new RBridge joins the campus and its originatingL1LSPBufferSize is smaller than current campus-wide Sz, reporting its originatingL1LSPBufferSize in its LSPs will cause other RBridges decrease their campus-wide Sz. Then the LSPs in the campus MUST be re-sized to be no greater than the new campus-wide Sz.

2) When an RBrige leaves the campus and its originatingL1LSPBufferSize is equal to the campus-wide Sz, its LSPs are purged from the remaining campus after reaching MaxAge [ISIS]. The campus-wide Sz MAY be recalculated and MAY increase. In other words, while RB1 normally ignores link state information for any IS-IS unreachable [RFC7180bis] RBridge RB2, originatingL1LSPBufferSize is an exception. Its value, even from IS-IS unreachable RBridges, is used in determining Sz.

Frequent LSP "re-sizing" is harmful to the stability of the TRILL campus, so it should be dampened. Within the two kinds of resizing actions, only the upward resizing will be dampened. When an upward resizing event happens, a timer is set (this is a configurable parameter whose default value is 300 seconds). Before this timer expires, all subsequent upward resizing will be dampened. Of course, in a well-configured campus with all RBridges configured to have the

same originatingL1LSPBufferSize, no resizing will be necessary. It does not matter if different RBridges have different dampening timers or some RBridges re-size upward more quickly than others.

If the refreshed campus-wide Sz is smaller than the lower bound or greater than the upper bound of the tested link MTU size, the resource consuming link MTU size testing can be avoided according to rule (a) or (b) specified in Section 3. Otherwise, RBridges need to test the link MTU size according to rule (c). But it's unnecessary to perform the link MTU size testing algorithm all over again.

5. Relationship between Port MTU, Lz and Sz

When port MTU size is smaller than the local originatingL1SNPBufferSize of an RBridge (sort of a wrong configuration), this port should be explicitly disabled from the TRILL campus. On the other hand, when an RBridge receives an LSP or E-L1CS FS-LSP with size greater than the link-wide Lz or the campus-wide Sz but not greater than its port MTU size, this LSP should be processed normally and not discarded. If the size of an LSP is greater than the MTU size of a port over which it is to be propagated, no attempt shall be made to propagate this LSP over the port and an LSPTooLargeToPropagate alarm shall be generated [ISIS].

6. LSP Synchronization

An RBridge participates in LSP synchronization on a link as soon as it has at least one adjacency on that link that has advanced to at least the 2-Way state [RFC7177]. On a LAN link, CSNP and PSNP PDUs are used for synchronization. On a point-to-point link, only PSNP are used.

The CSNPs and PSNPs MUST be formatted in chunks of size at most the link-wide Lz but are processed normally if received larger than that. Since the link MTU size may not have been tested in the 2-Way state, link-wide Lz may be greater than the supported link MTU size. In that case, a CSNP or PSNP may be discarded. After the link MTU size is successfully tested, RBridges will begin to format these PDUs in the size no greater than it, therefore these PDUs will eventually get through.

Note that the link MTU size is frequently greater than the campus-wide Sz. Link local PDUs are formatted in the size of link MTU size rather than the campus-wide Sz, which promises a reduction in the number of PDUs and a faster LSP synchronization process.

7. Recommendations for Traffic Link MTU Size Testing

Campus-wide Sz and link-wide Lz are used to limit the size of most TRILL IS-IS PDUs. They are different from the MTU size restricting the size of TRILL data packets. The size of a TRILL data packet is restricted by the physical MTU of the ports and links the packet traverses. It is possible that a TRILL data packet successfully gets through the campus but its size is greater than the campus-wide Sz or link-wide Lz values.

The algorithm defined in link MTU size testing can also be used in TRILL traffic MTU size testing; in that case the link-wide Lz used in that algorithm should be replaced by the port MTU of the RBridge sending MTU probes. The successfully tested size X can be advertised as an attribute of this link using MTU sub-TLV defined in [RFC7176].

Unlike RBridges, end stations do not participate in the exchange of ISIS PDUs of TRILL, therefore they cannot grasp the traffic link MTU size from a TRILL campus automatically. An operator may collect these values using network management tools such as TRILL ping or TraceRoute. Then the path MTU is set as the smallest tested link MTU on this path and end stations should not generate frames that, when encapsulated as TRILL Data packets, exceed this path MTU.

8. Backwards Compatibility

There can be a mixture of Lz-ignorant and Lz-aware RBridges on a link. This will act properly although it will not be as efficient as it would be if all RBridges on the link are Lz-aware.

At the side of an Lz-aware RBridge, in case that link-wide Lz is greater than campus-wide Sz, larger link-local TRILL IS-IS PDUs can be sent out to gain efficiencies. Lz-ignorant RBridges as receivers will have no problem handling them since the originatingL1LSPBufferSize value of these RBridges had been reported and the link-wide Lz is not greater than that value.

At the side of an Lz-ignorant RBridge, TRILL IS-IS PDUs are always formatted not greater than the campus-wide Sz. Lz-aware RBridges as receivers can handle these PDUs since they cannot be greater than the link-wide Lz.

An Lz-ignorant RBridge does not support the link MTU testing algorithm defined in Section 3 but may be using some algorithm just to test for Sz MTU on the link. In any case, if an RBridge per [RFC6325] receives an MTU-probe, it MUST respond with an MTU-ack padded to the same size as the MTU-probe. So the extension of TRILL MTU negotiation with Lz, as specified in this document, is fully backwards compatible.

9. Security Considerations

This document raises no new security issues for TRILL. For general and adjacency related TRILL security considerations, see [RFC6325] and [RFC7177].

10. IANA Considerations

IANA is requested to assign a new APPsub-TLV number from the range less than 256 in the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" registry for the TRILL originatingSNPBufferSize sub-TLV defined in Section 2 of this document. The entry is as follows:

Type	Name	Reference
tbd	originatingSNPBufferSize	[this document]

11. References

11.1. Normative References

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

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

[RFC7177] Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, May 2014.

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

[RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, September 2014.

[RFC7180bis] D. Eastlake, M. Zhang, et al., "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis, working in progress.

[RFC6823] Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, December 2012.

11.2. Informative References

[ISIS] ISO, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:2002.

[RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014.

Author's Addresses

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

EMail: zhangmingui@huawei.com

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

EMail: zhangxudong@huawei.com

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

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

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

EMail: radia@alum.mit.edu

Vishwas Manral
Ionos
4100 Moorpark Ave.
San Jose, CA 95117 USA

EMail: vishwas@ionosnetworks.com

Somnath Chatterjee
Cisco Systems

EMail: somnath.chatterjee01@gmail.com