INTERNET-DRAFT
Intended Status: Standards Track
Updates: 6325, 7177, 7780

M. Zhang
X. Zhang
D. Eastlake
Huawei
R. Perlman
EMC
S. Chatterjee
Cisco

Expires: February 16, 2017
August 15, 2016

Transparent Interconnection of Lots of Links (TRILL): MTU Negotiation
draft-ietf-trill-mtu-negotiation-05.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 state 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 recommended 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. This document updates RFC 6325, updates RFC 7177, and updates RFC 7780.

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 to fit in the campus-wide Sz.

[RFC7177] diagrams the state transitions of an adjacency. If MTU testing is enabled, "Link MTU size is successfully tested" is part of an event (event A6) causing the transition from "2-way" state to "Report" state for an adjacency. This 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. While in this document, a new RECOMMENDED link-wide minimum inter-RBridge MTU size, Lz, is specified. By calculating a using Lz as specified herein, link-scoped PDUs can be formatted greater than the campus-wide Sz up to the link-wide minimum acceptable inter-RBridge MTU size potentially improving the efficiency of link utilization and speeding link state convergence.

An optional TRILL MTU size testing algorithm is specified in Section 3 as an efficient method to update the old MTU testing method described in Section 4.3.2 of [RFC6325] and in [RFC7177]. The new MTU size testing method specified in this document is backward compatible to the old one. 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 MTU Size

This document specifies a new value "Lz" for the acceptable inter-RBridge link MTU size on a local link. Link-wide Lz is the minimum Lz supported between 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 are confined by the link-wide Lz instead of

the campus-wide Sz. CSNPs and PSNPs are examples of such PDUs. These PDUs are exchanged just on the local link. (While TRILL IS-IS Hellos are also link local, they are always limited to 1470 bytes for robustness.)

[RFC7356] defines the PDUs which support flooding scopes in addition to area-wide scope and domain-wide scope. As specified in [RFC6439bis], RBridges MUST support the Extended L1 Circuit-Scoped (E-L1CS) flooding scope LSP (FS-LSP). They use that flooding to exchange their maximally supportable value of "Lz". The smallest value of the Lz advertised by the RBridges on a link, but not less than Sz, is the link-wide Lz. An 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 [RFC7780], all RBridges MUST 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 TRILL GENINFO TLV [RFC7357] in an E-L1CS FS-LSP fragment zero. If it is missing from a fragment zero E-L1CS FS-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 FS-LSP fragment zero MUST NOT exceed 1470 bytes.

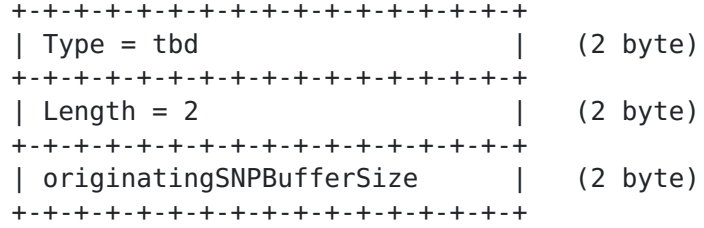

Figure 2.1: The originatingSNPBufferSize TLV.

Type: set to originatingSNPBufferSize APPsubTLV (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 as an unsigned integer, limited in the

range from 1470 to 65,535 bytes. (A value less than 1470 will be ignored.)

2.1. Operations

Lz is reported using a originatingSNPBufferSize TLV that MUST occur in fragment zero of the RBridge's E-L1CS FS-LSP. An originatingSNPBufferSize APPsub-TLV occurring in any other fragment is ignored. If more than one originatingSNPBufferSize APPsub-TLV occurs in fragment zero, the one advertising the smallest value for originatingSNPBufferSize, but not less than 1470 bytes, is used.

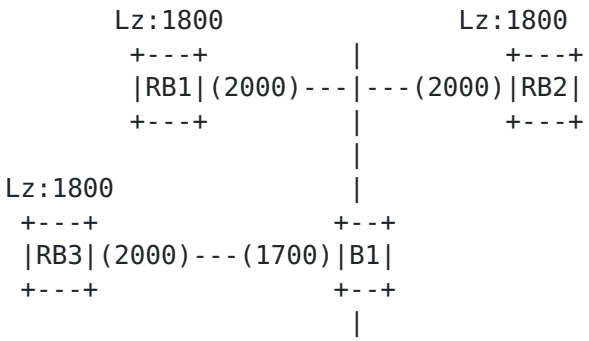

Figure 2.2: Link-wide Lz = 1800 v.s. tested link MTU size = 1700

Even if all RBridges on a specific link have reached consensus on the value of link-wide Lz based on advertised originatingSNPBufferSize, 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 CSNPs, PSNPs and E-L1CS FS-LSPs 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 formats their "campus-wide" PDUs, for example LSPs, not greater than what they determine as 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].

linkMtuSize, lowerBound, and upperBound 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), 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 is set to 1470, lowerBound is set to 1470, upperBound is set to the link-wide Lz, linkMtuSize is set to [(lowerBound + upperBound)/2] (Operation "[...]" returns the fraction-rounded-up integer.).

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

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

   upperBound is set to linkMtuSize and linkMtuSize is set to [(lowerBound + upperBound)/2]

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

   link MTU size is set to linkMtuSize, lowerBound is set to linkMtuSize and linkMtuSize is set to [(lowerBound + upperBound)/2]

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

4) 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 Designated RBRidge (DRB [RFC7177]) take the responsibility to do the testing. Multicast MTU-probes are used instead of unicast when multiple RBridges are desired to respond with an MTU-ack on the link. The Binary Search Algorithm given here is a way to minimize the probing attempts; 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 and lower bound 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 "lowerBound" >= campus-wide Sz. This link can support campus-wide Sz.

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

Otherwise, RBridges SHOULD test whether the link can support campus-wide Sz as in item (c) below. If they do not, the only safe assumption will be that the link cannot support Sz. This assumption, without testing, might rule out the use of a link that can, in fact, handle packets up to Sz. In the worst case, this might result in unnecessary network partition.

(c) "lowerBound" < campus-wide Sz < "upperBound". RBridges 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.

1) Joining

   a) 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 any LSP greater than the reduced Sz MUST be split and/or the LSP contents in the campus MUST be otherwise redistributed so that no LSP is greater than the new campus-wide Sz.

   b) If the joining RBridge's originatingL1LSPBufferSize is equal to or bigger than current campus-wide Sz, reporting its originatingL1LSPBufferSize will not change the campus-wide Sz.

2) Leaving

   a) From the specification of the Joining process, we know it's non-applicable that an RBridge leaves the campus while its origiatingL1LSPBufferSize is smaller than the campus-wide Sz.

   b) When an RBridge leaves the campus and its origiatingL1LSPBufferSize equals to the campus-wide Sz, its LSPs are purged from the remaining campus after reaching MaxAge [IS-IS]. The campus-wide Sz MAY be recalculated and MAY increase. In other words, while in most cases RB1 ignores link state information for IS-IS unreachable RBridge RB2 [RFC7780], originatingL1LSPBufferSize is meaningful. Its value, even from IS-IS unreachable RBridges, is used in determining Sz. This updates [RFC7780].

   c) When an RBrige leaves the campus and its originatingL1LSPBufferSize is greater than the campus-wide Sz, this will not update Sz since Sz is determined by another RBridge with smaller originatingL1LSPBufferSize.

Frequent LSP "re-sizing" is harmful to the stability of the TRILL campus, so, to avoid this, upward resizing SHOULD be dampened. When an upward resizing event is noticed by an RBridge, it is RECOMMENDED that a timer be set at that RBridge. This is a configurable parameter, LSPresizeTime, whose default value is 300 seconds. Before this timer expires, all subsequent upward resizing will be dampened (ignored). 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 test the link MTU size according to rule (c).

5. Relationship between Port MTU, Lz and Sz

When the port MTU of an RBridge is smaller than the local originatingL1SNPBufferSize of an RBridge (an inconsistent configuration), that port SHOULD be disabled and, in any case, an adjacency cannot be formed through such a port. 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 is processed normally. If the size of an LSP is greater than the MTU size of a port over which it is to be propagated, this LSP MUST NOT be sent over the port and an LSPTooLargeToPropagate alarm shall be generated [IS-IS].

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 that MTU, 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 limited in the size by the link MTU size rather than the campus-wide Sz, which, when Lz is greater than Sz, 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 for 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 is replaced by the port MTU of the RBridge sending MTU probes. The successfully tested size X MAY 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 TRILL IS-IS PDUs, 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 can be 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 may not be as efficient as it would be if all RBridges on the link are Lz-aware.

For 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.

For an Lz-aware RBridge, in the 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 tested and the link-wide Lz is not greater than that value.

An Lz-ignorant RBridge might not support the link MTU testing algorithm defined in Section 3 but could 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.

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. Additions to Configuration

Implementation of the features specified in this document adds two RBridge configuration parameters as follows:

10.1. Per RBridge Configuration

Each RBridge implementing the RECOMMENDED LSP re-sizing damping strategy specified in Section 4 has an LSPresizeTime parameter that is an integer in the range of 0-65,535 which defaults to 300. It is the number of seconds for which an RBridge determines that Sz has increased before it will create any LSP or E-L1FS FS-LSP fragments.

10.2. Per RBridge Port Configuration

Each RBridge port on which the calculation and use of Lz is implemented has an originatingL1SNPBufferSize parameter that is an integer in the range of 1,470-65,535. This parameter defaults to the minimum of the size that the port can accommodate and the size link-local IS-IS PDU that the TRILL implementation can accommodate.

11. 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]

12. Acknowledgements

Authors would like to thank the comments and suggestions from Vishwas Manral.

13. References

13.1. Normative References

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

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

Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <http://www.rfc-editor.org/info/rfc6325>.

[RFC7177] Eastlake 3rd, D., Perlman, R., Ghanwani, A., Yang, H., and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC 7177, DOI 10.17487/RFC7177, May 2014, <http://www.rfc-editor.org/info/rfc7177>.

[RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, DOI 10.17487/RFC7176, May 2014, <http://www.rfc-editor.org/info/rfc7176>.

[RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <http://www.rfc-editor.org/info/rfc7356>.

[RFC7780] Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <http://www.rfc-editor.org/info/rfc7780>.

[RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357, September 2014, <http://www.rfc-editor.org/info/rfc7357>.

13.2. Informative References

[IS-IS] International Organization for Standardization, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, November 2002.

[RFC6439bis] Eastlake 3rd, D., Yizhou, L., et al, "TRILL: Appointed Forwarders", draft-ietf-trill-rfc6439bis, Work in progress.

Author's Addresses

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

Phone: +86-13810702575
Email: zhangmingui@huawei.com

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

Email: zhangxudong@huawei.com

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

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

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

Email: radia@alum.mit.edu

Somnath Chatterjee
Cisco Systems
SEZ Unit, Cessna Business Park
Outer Ring Road
Bangalore - 560087
India

Email: somnath.chatterjee01@gmail.com