Link-State Routing
Internet-Draft
Intended status: Standards Track
Expires: April 15, 2019

H. Smit, Ed.

G. Van de Velde
Nokia

October 12, 2018

IS-IS Flooding over TCP
draft-hsmit-lsr-isis-flooding-over-tcp-00

Abstract

This document proposes a solution to use TCP for IS-IS flooding. The proposed solution is a relative simple extension to implement. IS-IS flooding over TCP brings BGP's property of scalable transport via TCP to Link-State protocols.

This proposal defines a new TLV in point-to-point IIHs to signal the intent of a router to do flooding over TCP, and it defines a small header to encapsulate IS-IS PDUs in the TCP byte-stream.

Requirements Language

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 [1].

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on April 15, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

IP Fabric Networks in data-centers are growing larger and larger. The number of routers in such fabrics are pushing the boundaries of routing protocols. Therefor new ideas are explored, and existing protocols are being enhanced (RFC 7938 [2], RIFT [5] and LSVR [4]).

This document improves an existing protocol, IS-IS. One of the scaling limitations of IS-IS is its flooding algorithm. BGP is known to be a routing protocol of high scale. A key property and important benefit of BGP is the fact that BGP uses TCP as transport mechanism. Introducing TCP to IS- IS flooding would bring a major positive scaling property from BGP to IS-IS.

2. General scaling factors of IS-IS

IS-IS is a highly scalable Interior Gateway Protocol (IGP). IS-IS is defined in ISO-10589 [3]. Networks with thousands of routers have been deployed. When bigger networks are build, certain parts of the algorithm become a limitation to the scalability of IS-IS.

Several sub-components of the IS-IS protocol have an impact on its scalability.

- The number of adjacenies. For each adjacency periodic IIHs have to be exchanged. Each adjacency has to be included in the router's Link-State PDU (LSP). When building a dynamic routing protocol, this work has to be done in some form or another. Not much can be done to improve the scalability of this effort.

- Flooding has a large impact on scalability of IS-IS. Obviously the number of LSPs in an area has an impact on the operation of IS-IS. Also the number of interfaces over which a router must flood has an impact on the operation of IS-IS. But the flooding algorithm itself has elements that limit scalability. Improving these sub-algorithms will have a positive impact on scalability.

- Dijkstra's Shortest-Path First algorithm. This algorithm is at the heart of Link-State protocols. This algorithm is computationally reasonably efficient. One could build better implementations, that do partial route-computation and do incremental SPF. Or that check the bi-directionality of each link in advance of running the SPF. One could run the regular SPF and the computations for LFA and rLFA in parallel. But the SPF algorithm itself can not be improved upon easily.

2.1. Current scaling limitations of IS-IS flooding

With current implementations of the IS-IS protocol, the flooding algorithms have the largest impact on protocol scalability. Three issues are particularly of concern.

2.1.1. Packet pacing and throughput

The first issue is packet pacing of LSPs. If routers would send large bursts of routing protocol packets to other routers, there is a risk that the receiving router might drop those packets. This risks increases when a router has multiple neighbors that might all be sending large amount of routing-protocol packets at the same time. Dropped packets cause re-transmissions, delays in convergence, or even worse things. The solution is packet pacing.

ISO-10589 suggests a router should wait 30 milliseconds between sending of two consecutive LSPs. This will give the receiving router time to process pending incoming packets, before input-queues get overwhelmed. This means that two routers can exchange at most 33 LSPs per second. If a router boots, and has an empty LSDB, in a network with 10000 routers (and thus at least 10000 LSPs), it will take up to 300 seconds before the new router has acquired the full LSDB.

Decreasing the inter-packet gap will speed this up, but it might have a negative impact on overall network stability. More dynamic or adaptive packet-pacing algorithms could be envisioned, but those are not public nor standardized. If such algorithms would be developed, they would probably end up including many aspects of the existing TCP protocol.

2.1.2. Reliable flooding on point-to-point interfaces

The second issue is implementing reliable flooding over point-to-point interfaces. The following algorithm is used when a LSP needs to be flooded:

- When a new LSP is received, the router sets the SRM-bits for this LSP for all interfaces (except the interface on which the new LSP was received).
- For each interface a pacing-timer is started (if not running yet).
- When that pacing-timer expires, the router will find an LSP with its SRM-bit set for that interface. It will transmit the LSP out over the interface.

- The router will then clear the SRM-bit for that LSP. It will set a bit indicating that this LSP has not been acknowledged yet. And it will start a retransmit-timer for that LSP/interface combination.

- When a PSNP is received for this LSP on this interface, the router will clear the bit that indicates that no acknowledgement was received yet.

- When the retransmit-timer fires, the router will check whether the retransmit-it has been cleared yet. If so, the router stops the retransmit-timer and is done. If the retransmit-bit has not been cleared, then the router sets the SRM-bit for this interface/LSP combination again, and start the pacing-timer for the interface (if not still running).

Note that when flooding LSPs, the router needs to keep a retransmit-timer per LSP/interface combination. These timers run typically for 5 seconds, or until an acknowledgement (PSNP) is received. In a network with only a few hunderd LSPs, then 5 seconds * 33 LSPs/second results in only 165 LSPs being flooded. If the router has 100 interfaces, this can cause the router to have 16500 simultanously running timers. If a router falls behind processing PSNPs, or when PSNPs are being dropped, this number could increase to even larger numbers. The conclusion is that reliability of flooding LSPs over point-to-point interfaces does not come free. And in networks under stress, the cost can become even higher.

2.1.3. Unreliability of CSNPs

The third issue of concern is the unreliability of CNSPs. CSNPs are used when flooding over multi-point interfaces. But CSNPs are also used to synchronize LSDBs over adjacencies on point-to-point interfaces. This happens right after a new adjacency over a point-to-point interface is established. The algorithm used after a new adjacency comes up is:

- The router sets the SRM-bit for this interface on all LSPs in its LSDB.

- The router creates CSNPs describing all LSPs in its LSDB. It sends these CSNPs to the new neighbor.

- The router waits a limited amount of time, hoping to receive all the CSNPs from its new neighbors.

- For every LSP in every CSNP received from its new neighbor, the router checks to compare its version of the LSP with its neighbors

version of the LSP. If the versions are the same, the router clears the SRM-bit for that LSP/interface. Versions are compared using the LSP sequence-number (and checksum, TTL, etc).

- The router starts the packet-pacing timer, and starts sending to the new neighbor, LSPs that still have the SRM-bit set for that interface.

When the number of LSPs in the LSDB grows to large numbers, the number of CSNPs needed increases to large numbers as well. There can be only descriptions of 91 LSPs in a typical CSNP. If a network has 10000 routers, and thus 10000+ LSPs, it takes 110 CNSPs to describe the whole LSDB. If any of the CSNPs that get exchanged during adjacency synchronization gets dropped, the sending router will transmit 91 LSPs per dropped CNSPs, regardless whether that was necessary or not.

3. Improvements for IS-IS flooding

BGP is considered to be a highly scaleable routing protocol. It is used to carry all routes in the Global Internet. It is used to carry large numbers of customer routes in Provider networks that supply VPN-services. But BGP has downsides too. BGP typically requires extra configuration, and in dense topologies routing-churn can be experienced, because BGP does so-called path-hunting.

The main property of BGP that contributes to good scaling is the fact that BGP uses TCP for its transport. Using TCP has certain benefits for a routing protocol. TCP supplies reliability through retransmissions and acknowledgements. TCP supplies high throughput through its windowing mechanism and by potentially packing small chunks of user-data into larger TCP segments. TCP supplies a crude form of multi-threading by separating transmission and retransmission of data from the user process, and letting other tasks or the kernel take care of that. When a routing protocol uses TCP, it does not need to burden itself anymore with tasks like retransmission, acknowledgements, flow-control, or seeking high bandwidth and throughput. It also doesn't need to do extras to use multi-threading for reliable transmission.

3.1. Using TCP to do IS-IS flooding

This document proposes a relatively simple way to do IS-IS flooding over TCP.

Routers remain to establish new adjacencies using IIHs via the classic IS-IS mechanism. When using IS-IS TCP extensions Routers remain sending periodic IIHs via the classic mechanism to maintain

adjacencies. However, after establishing a new adjacency and successfully establish a corresponding TCP-connection, LSPs and SNPs are sent only over the TCP-connection.

4. Negotiating Flooding over TCP

Before two routers can start flooding over TCP, they need to agree on this new way of transport. Negotiating is done via a new TLV in the IS-to-IS Hello PDUs (IIH). When a router intends to do flooding over TCP, it includes this new TLV in its p2p IIHs. The existence of the TLV in the IIHs is an indication to the other router that it wants to use TCP for flooding.

The size of the TLV is variable. The value field contains the IP address and TCP port-number on which the router is accepting a new TCP connection for flooding. The router with the lowest System-ID initiates the TCP connection to the other router. The router with the highest System-ID never tries to set up a new connection. It just listens on its advertised TCP port-number and accepts the TCP connection from the router with the lower System-ID.

If only one, or neither of both routers include the new TLV in their IIHs, then flooding will not be done over TCP. Instead the classic IS-IS flooding algorithm is used, as described in ISO-10589.

Flooding over TCP is only supported on point-2-point interfaces.

4.1. The new TLV to indicate that a router wants to flood over TCP

This document proposes a new TLV for IIH messages. The existence of this TLV in an IIH signals the receiving router that the sending router is willing to do flooding over TCP.

The content of the TLV are 2 or more sub-TLVs. These sub-TLVs indicate the TCP port-number on which the advertising router is listening to accept new TCP-connections, and 1 or more sub-TLVs that indicate the IPv4- or IPv6-address on which the router is listening to accept new TCP-connections.

The new TLV consists of:

- 1 octet of TLV-Type,
- 1 octet of TLV-Length,
- 10 to 255 octets of TLV-Value, containing sub-TLVs.

The defined sub-TLVs are:

- TLV type 1, Length 2 octets. TCP port number.

- TLV type 2, Length 4 octets. IPv4 address. The sending router is listening on this IPv4-address to open a TCP-connection for IS-IS flooding.

- TLV type 3, Length 16 octets. IPv6 address. The sending router is listening on this IPv6-address to open a TCP-connection for IS-IS flooding.

Example of the layout of the new Flooding-over-TCP TLV. This example advertises an IPv4-address to connect to.

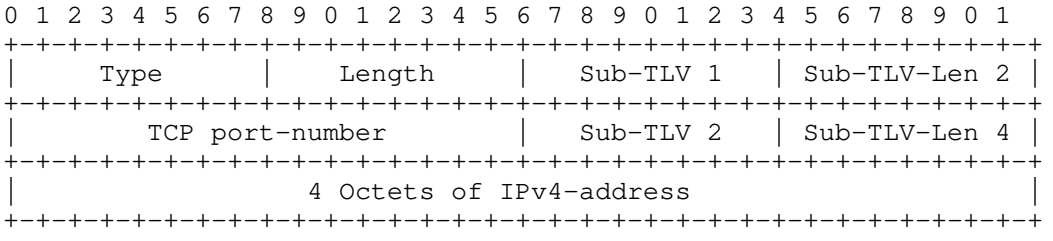

Figure 1

The new TLV is to be included only once in each IIH. A router MUST NOT include more than one TCP port number sub-TLV. A router MAY include multiple IPv4- or IPv6-address sub-TLVs. The destination IP-address(es) SHOULD be addresses that are also included in the IP-Interface-Addresses TLVs (TLV 132 for IPv4 or TLV 233 for IPv6).

5. Format of messages over TCP

The content of the messages that are transmitted over the TCP connection are traditional IS-IS PDUs. IIHs, SNPs and LSPs can all be transmitted over the TCP connection. No TLV-format or other extensible format is needed, because new information is to be included inside IIHs, SNPs or LSPs themselves. Therefor the format of messages over TCP itself does not need to be changed, and does not need to be extensible.

Each IS-IS PDU that is sent over TCP is to be preceeded by a header, functionaing as a marker. This header consists of:

- Four octets of marker. The content of this marker is always 0x69 0x73 0x69 0x73. This marker has the same function as the marker

in a BGP-header. It enables the receiver to check whether messages inside the TCP-bytestream have gone out of sync.

- Two octets of message-length. The IS-IS PDU itself also has a length-field, inside the message-specific header. The length-field here can be used to verify no octets are missing and that there are no extra trailing octets.

The type of IS-IS PDU can be derived from the PDU itself, by looking at the "PDU Type" field in the common IS-IS PDU header.

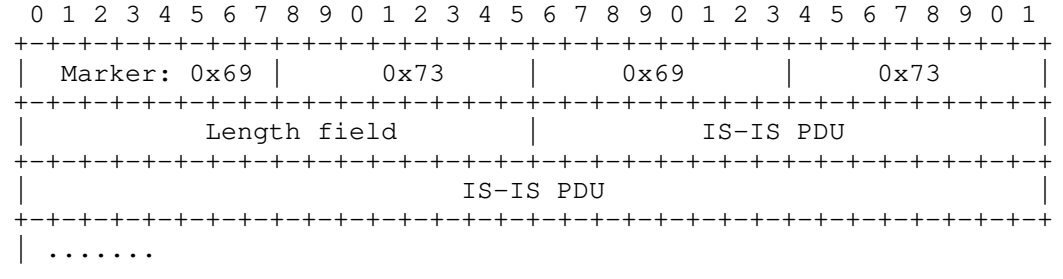

Figure 2

6. New behaviour of IS-IS flooding

When IS-IS does its flooding over TCP, the algorithms to transmit and receive LSPs change slightly. The biggest difference with the standard algorithms from ISO-10589 are the facts that the sending router does not need to do retransmission and that the receiving router does not need to send PSNPs to acknowledge receipt of LSPs.

6.1. Establishing a new IS-IS adjacency

Initially the Router looks for the new TLV in the IIH. If the other router included this TLV in its IIH, flooding over TCP is initiated. The router with the lowest System-ID initiates a TCP-connection to the other router. The TCP port-number and destination IP-address is learned from the new TLV in the IIH.

After the TCP session is established, a router will:

- Send a regular IIH over the TCP connection. The IIH is the same as when it would have been encapsulated straight into a layer-2 header. This IIH allows the other router to verify the identity and authentication of the remote router.

- Wait for receipt of an IIH from the remote router. This IIH is used to verify the identity and authentication of the remote router.

- Set the SRM-bit for this interface on all the LSPs in the LSDB.

- Send a number of CSNPs over the TCP connection. These CNSPs MUST describe the whole LSDB of the sending router. The last CSNP should desribe the last lexographical LSP in the LSDB. The end-id in the CSNP would be FFFF.FFFF.FFFF.FF-FF.

- Process all incoming CSNPs from the remote router. When a CSNP is received, check your own LSDB, and clear the SRM-bits on LSPs that both routers have in common. If the remote router has a version of the LSP that is newer, do not set the SSN-bit. It is not necessary to explicitely request for the newer LSP. The remote router will send it anyway.

- When the last CSNP has been received, walk the LSDB and send any LSPs that still have the SRM-bit set for this interface.

- No retransmission needs to be one by either router. TCP will take care of retransmission.

6.2. Behaviour during the existence of an IS-IS adjacency

The actions that a router has to take when receiving a new LSP are simplified compared to classic flooding.

- When a router receives an LSP, it checks if it has that LSP already in its LSDB. And it checks whether the version of the received LSP is newer or not.

- If the version is the same, the router does nothing.

- If the version of the received LSP is older than the LSP in the LSDB, the router sets the SRM-bit for the LSP. At some point in time, the router will then send its own LSP back to the other router.

- If the version of the recieved LSP is newer than the LSP in the LSDB, the router sets the SRM-bits for this LSP for all interfaces, except the interface it received the newer version of the LSP from.

- The receiving router does not set the SSN-bit and does not send an acknowledgement (PSNP).

- Periodically, or event driven, the router will check its LSDB for LSPs with the SRM-bit set. When it finds such LSPs, it will send as many of those LSPs to neighbors, via TCP. There is no packet-pacing. All flow-control is handled by TCP. After sending one or more LSPs, the router does not set any state to indicate that the LSP needs retransmission. The router does not expect an acknowledgement (PSNP). No retransmission-timer needs to be started. Just sending the LSPs is enough.

7. Considerations regarding IS-IS flooding over TCP

7.1. Flooding over TCP is only done on point-to-point interfaces

Flooding over TCP is not supported for multi-point interfaces. The nature of classic flooding between multiple routers on a LAN is too complex to simply replace by TCP connections. Therefor the new flooding-over-TCP TLV should only be included in point-to-point IIH.

Care must be taken that when a large network consists mostly of point-to-point interfaces, there are no multi-point between routers left in the network. Doing classic flooding over those multi-point interfaces might require substantial more resources than flooding on routers with only point-to-point interfaces.

7.2. Unnumbered interfaces and reachability of the interface ip-address

When a router tries to open a TCP connection to another router, it uses the TCP port-number and an IP-address is has learned from the new flooding-over-TCP TLV. This destination address can be any advertised IP-address that is from a prefix shared between the two routers.

However, it is possible that both routers use "ip unnumbered" on the point-to-point interface. In that case, the remote destination ip-address might not appear in the sender's routing table. Typically routes are installed in the routing table only after doing an SPF computation and learning how to reach all IP-prefixes that are included in LSPs. Typically routers do not install routes in the routing table for IP-addresses learned from the IP-Interface-Addresses TLV in IIHs. When a router is planning to do flooding over TCP, and does not have opened a TCP connection yet, it will not have all the LSPs in its LSDB necessary to learn how to reach the IP-address from the new Flooding-over-TCP TLV, or from the IP-Interface-Addresses TVL.

Therefor it is recommended that when a router does flooding over TCP, and one of its interfaces is configured as "unnumbered", that router SHOULD install host-routes (/32s or /128s) in its routing table, so

that TCP will be able to open a connection to the router on the other end of an adjacency. These host-routes can be interface-routes for the IP-address(es) learned from the new Flooding-over-TCP TLV in the p2p IIHs.

7.3. Multiple levels of hierarchy on one interface

IS-IS flooding over TCP is only defined for point-to-point interfaces. Over point-to-point interfaces, only one type of IIH PDU is sent, even when the interface is used by both level-1 and level-2 routing. This means that IS-IS flooding over TCP is negotiated in only one location (inside the p2p IIH). Two routers use a single TCP-connection, even when doing both level-1 and level-2 routing over that interface.

The packet-types of LSPs and SNPs identify whether the packet is level-1 or level-2. Therefor no confusion can occur when receiving both level-1 and level-2 PDUs over the same TCP connection.

7.4. Downsides of using TCP for IS-IS flooding

When TCP-segments are dropped, TCP will retransmit those segments a little while later. In the mean time, new versions might arrive of the LSPs that are in the TCP buffers. Therefor TCP might retransmit stale LSPs. Which it would not have done if flooding was done via the standard way. This causes only a slight inefficient use of resources. Ultimately the current versions of those LSPs will be transmitted. To protect against this, it is recommended to not make the TCP window-size larger than the default.

7.5. What to do if the TCP connection breaks

If a TCP connection gets closed or reset, the router with the lowest System-ID MUST periodically try to re-open the TCP connection. Both routers MUST NOT declare the adjacency down. An existing adjacency must stay established as long as IIHs are exchanged and the holding-time timer doesn’t expire.

The benefit of this behaviour is that it allows IS-IS implementations a certain flexibility. E.g. when an IS-IS process on a router is restarted, and the TCP connection is re-established, this will not bring down the adjacency. Or a router can switch over to the Hot Standby Control Plane, or do In-Service Software-Upgrades (ISSU) without causing adjacencies to go down.

7.6. What to do if the TCP connection can not be set up

It could happen that two routers can exchange IS-IS PDUs fine, but they can not set up a TCP connection. What needs to be done in this case is open for discussion.

8. Security Considerations

IS-IS as defined in ISO-10589 encapsulates IS-IS PDUs straight into a layer-2 header. One of the benefits of this is that remote attackers can not send IS-IS messages to a targeted router that is several ip-hops away. Using TCP for IS-IS flooding would potentially open up IS-IS routers to these forms of attacks.

The common way for a protocol to protect itself against these remote attack is using the TTL-field in the IP-header of TCP-segments.

When a router send a TCP-segment with IS-IS flooding data, it MUST set the TTL of the IP-header to 255. When a router receives a TCP-segment with IS-IS flooding data, it MUST check to see if the TTL is still 255. If a router receives a TCP-segment with IS-IS flooding data, and the TTL is less than 255, the router MUST ignore and drop the TCP-segment.

Identification and Authentication. When a new TCP-session is established to flood over, each router MUST first send a regular IIH over the TCP-session. This allows each router to verify that the other side of the TCP-connection is who they expect it to be. The IIH has the System-ID and the Interface-ID of the sending router. Regular authentication methods will place an authentication-TLV inside the IIH. Regardless of the fact whether routers flood over layer-2 or flood over TCP, these authentication mechanisms can be used to verify the other side of the TCP-connection. Sending a regular IIH for verification and authentication, instead of having our own new method, guaranteees that Flooding-over-TCP will use new authentication mechanisms when those get developed in the future.

9. Acknowledgements

The authors would like to express thanks to Filip Martin, Dirk Goethals and Koen Leclerq for their suggestions and comments.

10. Contributor Addresses

Below is a list of other contributing authors in alphabetical order:

Figure 3

11. IANA Considerations

This document requests one new TLV code-point, to be used in IIHs

12. References

12.1. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://xml.resource.org/public/rfc/html/rfc2119.html>.

[2] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

12.2. Informative References

[3] International Standard 10589, "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), Second Edition.", 2002.

[4] Patel, K., Lindem, A., and W. Henderickx, "Link State Vector Routing (LSVR)", August 2018.

[5] Przygienda, T., Sharma, A., Thubert, P., Atlas, A., and J. Drake, "Routing in Fat Trees (RIFT)", June 2018.

Authors' Addresses

Henk Smit (editor)

Email: hhw.smit@xs4all.nl

Gunter Van de Velde
Nokia
Copernicuslaan 50
2018 Antwerp
Belgium

Email: gunter.van_de_velde@nokia.com