Workgroup: Network Working Group
Internet-Draft: draft-ietf-lsr-distoptflood-02
Published: 7 August 2023
Intended Status: Experimental
Expires: 8 February 2024
Authors: R. White (Akamai)
S. Hegde (Juniper Networks)
T. Przygienda (Juniper Networks)

IS-IS Optimal Distributed Flooding for Dense Topologies

Abstract

In dense topologies (such as data center fabrics based on the Clos and butterfly topologies, though not limited to those exclusively), IGP flooding mechanisms designed originally for sparse topologies can "overflood," or in other words generate too many identical copies of topology and reachability information arriving at a given node from other devices. This normally results in slower convergence times and higher resource utilization to process and discard the superfluous copies. The modifications to the flooding mechanism in the Intermediate System to Intermediate System (IS-IS) link state protocol described in this document reduce resource utilization significantly, while increaseing convergence performance in dense topologies. Beside reducing the extraneous copies it uses the dense topologies to "load-balance" flooding across different possible paths in the network to prevent build up of flooding hot-spots.

Note that a Clos fabric is used as the primary example of a dense flooding topology throughout this document. However, the flooding optimizations described in this document apply to any arbitrary topology.

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 8 February 2024.

Table of Contents

1. Introduction

1.1. Goals

The goal of this draft is to solve one of the problems occurring when operating a link state protocol in a densely meshed topology. Such topologies with high average fanout, causes too many copies of identical information to be flooded within the network. Analysis and experiments show, for instance, that in a butterfly fabric of around 2'500 intermediate systems, each intermediate system will receive over 40 copies of any changed LSP fragment. This not only wastes bandwidth and processor time, this dramatically slows convergence speed under topological changes.

This document describes a set of modifications to the existing IS-IS flooding mechanisms which will minimize the number of LSP fragments received by individual intermediate systems. In its extreme version the change leads to only one copy per intermediate system being processed. The mechanisms described in this document are similar to and based on those implemented in OSPF to support mobile ad-hoc networks, as described in[RFC5449],[RFC5614]. These solutions have been widely implemented and deployed.

1.2. Contributors

The following people have contributed to this draft and are mentioned without any particular order: Abhishek Kumar, Nikos Triantafillis, Ivan Pepelnjak, Christian Franke, Hannes Gredler, Les Ginsberg, Naiming Shen, Uma Chunduri, Nick Russo, and Rodny Molina.

1.3. Experimental Evidence

Laboratory tests based on a well known open source codebase show that modifications similar to the ones described in this draft reduce flooding in a large scale emulated butterfly network topology signficantly. Under unmodified flooding procedurs intermediate systems receive, on average, 40 copies of any changed LSP fragment in a 2'500 nodes butterfly network. With the changes described in this document said systems received, on average, two copies of any changed LSP fragment. In many cases, only a single copy of each changed LSP was received and processed per node. In terms of performance, overall convergence times were cut in roughly half.

An early version of mechanisms described in this document has been implemented in the FR Routing open source routing stack as part of `fabricd` daemon.

1.4. Example Network

Following spine and leaf fabric will be used in further description of the introduced modifications.

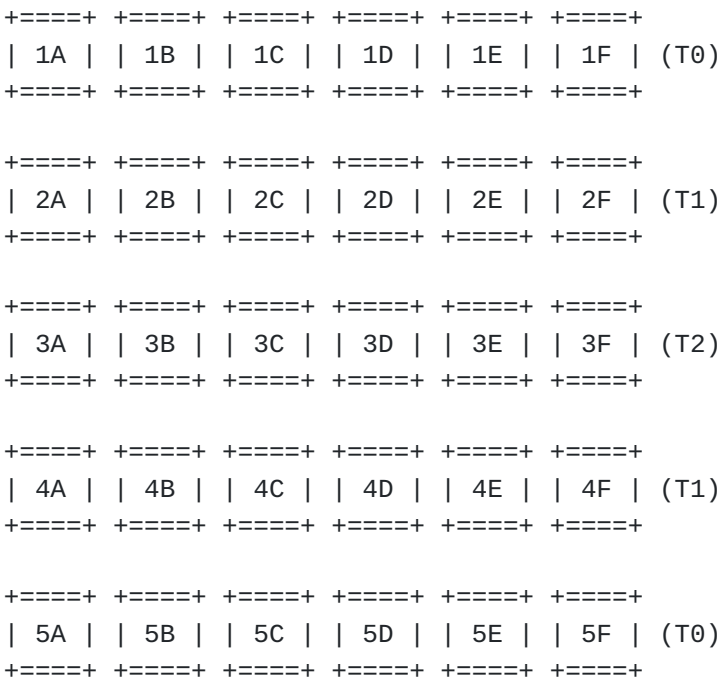

Figure 1

The above picture does not contain the connections between devices for readability purposes. The reader should assume that each device in a given layer is connected to every device in the layer above it in a butterfly network fashion. For instance:

- *5A is connected to 4A, 4B, 4C, 4D, 4E, and 4F
- *5B is connected to 4A, 4B, 4C, 4D, 4E, and 4F
- *4A is connected to 3A, 3B, 3C, 3D, 3E, 3F, 5A, 5B, 5C, 5D, 5E, and 5F
- *4B is connected to 3A, 3B, 3C, 3D, 3E, 3F, 5A, 5B, 5C, 5D, 5E, and 5F
- *etc.

The tiers or stages of the fabric are marked for easier reference. Alternate representation of this topology is a "folded Clos" with T2 being the "top of the fabric" and T0 representing the leaves.

2. Flooding Modifications

This section describes detailed modifications to the IS-IS flooding process to reduce flooding load in a densely meshed topology. It does at the same time distribute the reduced flooding across the whole topology to prevent hot-spots.

2.1. Optimizing Flooding

The simplest way to conceive of the solution presented here is in two stages:

- *Stage 1: Forward Optimization
  - -Find the group of intermediate systems that will all flood to the same set of neighbors as the local IS
  - -Decide (deterministically) which subset of the intermediate systems within this group should re-flood any received LSPs
- *Stage 2: Reverse Optimization
  - -Find neighbors on the shortest path towards the origin of the change
  - -Do not flood towards these neighbors

The first stage is best explained through an illustration. In the network above, if 5A transmits a modified Link State Protocol Data Unit (LSP) to 4A-4F, each of 4A-4F nodes will, in turn, flood this modified LSP to 3A (for instance). With this, 3A will receive 6 copies of the modified LSP, while only one copy is necessary for the intermediate systems shown to converge on the same view of the topology. If 4A-4F could determine that all of them will all flood identical copies of the modified LSP to 3A, it would be possible for all of them except one to decide not to flood the changed LSP to 3A.

The technique used in this draft to determine such flooding group is for each intermediate system to calculate a special SPT (shortest-path spanning tree) from the point of view of the transmitting neighbor. As next step, by setting the metric of all links to 1 and truncating the SPT to two hops, the local IS can find the group of neighbors it will flood any changed LSP towards and the set of intermediate systems (not necessarily neighbors) which will also flood to this same set of neighbors. If every intermediate system in the flooding set performs this same calculation, they will all obtain the same flooding group.

Once such a flooding group is determined, the members of the flooding group will each (independently) choose which of the members should re-flood the received information. A common hash function is used across a set of shared variables so each member of the group comes to the same conclusion as to the designated flooding nodes. The group member which is in such a way `selected` to flood the changed LSP does so normally; the remaining group members suppress the flooding of the LSP initially.

Note that there is no signaling between the intermediate systems running this flooding reduction mechanism for the solution to work. Each IS calculates the special, truncated SPT separately, and determines which IS should flood any changed LSPs independently based on a common hash function. Because these calculations are performed using a shared view of the network, however (based on the common link state database) and such a shared hash function, each member of the flooding group will make the same decision under converged conditions. In the transitory state of nodes having potentially different view of topologies the flooding may either overflood or in worse case not flood enough for which we introduce a 'quick-patching' mechanism later but ultimately will converge due to periodic CSNP origination per normal protocol operation.

The second stage is simpler, consisting of a single rule: do not flood modified LSPs along the shortest path towards the origin of the modified LSP. This rule relies on the observation that any IS between the origin of the modified LSP and the local IS should receive the modified LSP from some other IS closer to the source of the modified LSP. It is worth to observe that if all the nodes that should be designated to flood within a peer group are pruned by the second stage the receiving node is at the `tail-end` of the flooding chain and no further flooding will be necessary. Also, per normal protocol procedures flooding to the node from which the LSP has been received will not be performed.

2.2. Optimization Process Details

This section provides normative description of the specification. Any node implementing this solution MUST exhibit external behavior that conforms to the algorithms provided.

Each intermediate system will determine whether it should re-flood LSPs as described below. When a modified LSP arrives from a Transmitting Neighbor (TN), the result of the following algorithm obtains the necessary decision:

Step 1: Build the Two-Hop List (THL) and Remote Neighbor's List (RNL) by:

A) Set all link metrics to 1

B) Calculate an SPT truncated to 2 hops from the perspective of TN

C) For each IS that is two hops away (has a metric of two in the truncated SPT) from TN:

i. If the IS is in a neighbor of the LSP originator, skip

ii. If the IS is on the shortest path towards the originator of the modified LSP, skip

iii. If the IS is *not* on the shortest path towards the originator of the modified LSP, add it to THL

D) Add each IS that is one hop away from TN to the RNL

Step 2: Sort nodes in RNL by system IDs, from the least value to the greatest.

Step 3: Calculate a number, N, by adding first each byte in LSP-ID under consideration (without using the fragment ID) and then adding value of its fragment ID MOD 2 (footnote 1: this allows for some balancing of LSPs coming from same system ID without introducing excessive amount of state in an implementation per originator). Consequently, set N to the MOD of N when divided by number of neighbors in RNL. With that N will be less than the number of members of RNL.

Step 4: Starting with the Nth member of RNL:

A) If THL is empty, exit

B) If this member of RNL is the local calculating IS, it MUST reflood the modified LSP; exit

C) Remove all members of THL connected to (adjacent to) this member of RNL

D) Move to the next member of RNL, wrapping to the beginning of RNL if necessary

Note 1: This description is leaning towards clarity rather than optimal performance when implemented.

Note 2: An implementation in a node MAY choose independently of others to provide a configurable parameter to allow for more than one node in RNL to reflood, e.g. it may reflood even if it's only the member that would be chosen from the RNL if a double coverage of THL is required. The modifications to the algorithm are simple enough to not require further text.

2.3. Flooding Failures

It is possible that during initial convergence or in some failure modes the flooding will be incomplete due to the optimizations outlined. Specifically, if a reflooder fails, or is somehow disconnected from all the links across which it should be

reflooding, an LSP could be only partially distributed through the topology. To speed up convergence under such partition failures (observe that periodic CSNPs will under any circumstances converge the topology though at a slower pace), an intermediate system which does not reflood a specific LSP (or fragment) SHOULD:

A) Set a short, configurable timer which should be significantly shorter than CSNP interval used.

B) When the timer expires, send Partial Sequence Number Packet (PSNP) of all LSPs that have *not* been reflooded during the timer runtime to all neighbors unless an up-to-date PSNP or CSNP has been already received from the neighbor.

C) Per normal protocol procedures process any Partial Sequence Number Packets (PSNPs) received that indicate that neighbors still have older versions of the LSP will lead to the usual synchronization of the databases that are out of sync due to optimized flooding.

D) If such resynchronizations above a configurable threshold are required (i.e. PSNPs are sent to the neighbors and are answered with requests), an implementation SHOULD notify the network operator via the according mechanism about the condition.

2.4. Signaling

A node deploying this algorithm SHOULD advertise algorithm value <TBD> in the IS-IS Dynamic Flooding sub-TLV of the Router Capability TLV (242) [RFC7981] as specified in [I-D.ietf-lsr-dynamic-flooding]. It bares repeating again that in case the hashing algorithm a node uses is different from this draft a different algorithm number must be assigned and used.

2.5. Additional Deployment Considerations

A node deploying this algorithm on point-to-point links MUST send CSNPs on such links. This does not represent a dramatic change given most deployed implementations today already exhibit this behavior to prevent possible slow synchronization of IS-IS database across such links and to provide additional periodic consistency guarantees.

2.6. Flooding Example

Assume, in the network specified, that 5A floods some modified LSP towards 4A-4F and we only use a single node to reflood. To determine whether 4A should flood this LSP to 3A-3F:

* 5A is TN; 4A calculates a truncated SPT from 5A's perspective with all link metrics set to 1

* 4A builds THL, which contains 3A, 3B, 3C, 3D, 3E, 3F, 5B, 5C, 5D, 5E and 5F

* 4A builds RNL, which contains 4A,4B,4C,4D,4E and 4F, sorting it by the system ID

* 4A computes hash on the received LSP-ID to get N; assume N is 1 in this case

* Since 4A is the 1st member of RNL and there are members in THL, 4A must reflood; the loop exits

2.7. A Note on Performance

The calculations described here seem complex, which might lead the reader to conclude that the cost of calculation is so much higher than the cost of flooding that this optimization is counter-productive. First, The description provided here is designed for clarity rather than optimal calculation. Second, many of the involved calculations can be easily performed in advance and stored, rather than being performed for each LSP occurence and each neighbor. Optimized versions of the process described here have been implemented, and do result in strong convergence speed gains.

3. Security Considerations

This document outlines modifications to the IS-IS protocol for operation on high density network topologies. Implementations SHOULD implement IS-IS cryptographic authentication, as described in [RFC5304], and should enable other security measures in accordance with best common practices for the IS-IS protocol.

4. References

4.1. Normative References

[I-D.ietf-lsr-dynamic-flooding] Li, T., Psenak, P., Chen, H., Jalil, L., and S. Dontula, "Dynamic Flooding on Dense Graphs", Work in Progress, Internet-Draft, draft-ietf-lsr-dynamic-flooding-14, 8 June 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-lsr-dynamic-flooding-14>.

[ISO10589] ISO, "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.

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

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <https://www.rfc-editor.org/info/rfc2629>.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, DOI 10.17487/RFC5301, October 2008, <https://www.rfc-editor.org/info/rfc5301>.

[RFC5303] Katz, D., Saluja, R., and D. Eastlake 3rd, "Three-Way Handshake for IS-IS Point-to-Point Adjacencies", RFC 5303, DOI 10.17487/RFC5303, October 2008, <https://www.rfc-editor.org/info/rfc5303>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, DOI 10.17487/RFC5308, October 2008, <https://www.rfc-editor.org/info/rfc5308>.

[RFC5309] Shen, N., Ed. and A. Zinin, Ed., "Point-to-Point Operation over LAN in Link State Routing Protocols", RFC 5309, DOI 10.17487/RFC5309, October 2008, <https://www.rfc-editor.org/info/rfc5309>.

[RFC5311] McPherson, D., Ed., Ginsberg, L., Previdi, S., and M. Shand, "Simplified Extension of Link State PDU (LSP)

Space for IS-IS", RFC 5311, DOI 10.17487/RFC5311, February 2009, <https://www.rfc-editor.org/info/rfc5311>.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <https://www.rfc-editor.org/info/rfc5316>.

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

[RFC7981] Ginsberg, L., Previdi, S., and M. Chen, "IS-IS Extensions for Advertising Router Information", RFC 7981, DOI 10.17487/RFC7981, October 2016, <https://www.rfc-editor.org/info/rfc7981>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

4.2. Informative References

[I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", Work in Progress, Internet-Draft, draft-ietf-isis-segment-routing-extensions-25, 19 May 2019, <https://datatracker.ietf.org/doc/html/draft-ietf-isis-segment-routing-extensions-25>.

[RFC3277] McPherson, D., "Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance", RFC 3277, DOI 10.17487/RFC3277, April 2002, <https://www.rfc-editor.org/info/rfc3277>.

[RFC3719] Parker, J., Ed., "Recommendations for Interoperable Networks using Intermediate System to Intermediate System (IS-IS)", RFC 3719, DOI 10.17487/RFC3719, February 2004, <https://www.rfc-editor.org/info/rfc3719>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <https://www.rfc-editor.org/info/rfc5304>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC5449] Baccelli, E., Jacquet, P., Nguyen, D., and T. Clausen, "OSPF Multipoint Relay (MPR) Extension for Ad Hoc Networks", RFC 5449, DOI 10.17487/RFC5449, February 2009, <https://www.rfc-editor.org/info/rfc5449>.

[RFC5614] Ogier, R. and P. Spagnolo, "Mobile Ad Hoc Network (MANET) Extension of OSPF Using Connected Dominating Set (CDS) Flooding", RFC 5614, DOI 10.17487/RFC5614, August 2009, <https://www.rfc-editor.org/info/rfc5614>.

[RFC5820] Roy, A., Ed. and M. Chandra, Ed., "Extensions to OSPF to Support Mobile Ad Hoc Networking", RFC 5820, DOI 10.17487/RFC5820, March 2010, <https://www.rfc-editor.org/info/rfc5820>.

[RFC5837] Atlas, A., Ed., Bonica, R., Ed., Pignataro, C., Ed., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, DOI 10.17487/RFC5837, April 2010, <https://www.rfc-editor.org/info/rfc5837>.

[RFC6232] Wei, F., Qin, Y., Li, Z., Li, T., and J. Dong, "Purge Originator Identification TLV for IS-IS", RFC 6232, DOI 10.17487/RFC6232, May 2011, <https://www.rfc-editor.org/info/rfc6232>.

[RFC7921] Atlas, A., Halpern, J., Hares, S., Ward, D., and T. Nadeau, "An Architecture for the Interface to the Routing System", RFC 7921, DOI 10.17487/RFC7921, June 2016, <https://www.rfc-editor.org/info/rfc7921>.

Authors' Addresses

Russ White
Akamai

Email: russ@riw.us

Shraddha Hegde
Juniper Networks

Email: shraddha@juniper.net

Tony Przygienda

Juniper Networks

Email: prz@juniper.net