

Networking Working Group
Internet-Draft
Intended status: Informational
Expires: January 9, 2022

L. Ginsberg
P. Psenak
M. Karasek
A. Lindem
Cisco Systems
T. Przygienda
Juniper
July 8, 2021

IS-IS Flooding Scale Considerations
draft-ginsberg-lsr-isis-flooding-scale-05

Abstract

Link State PDU flooding rates in use are much slower than what modern networks can support. The use of IS-IS at larger scale requires faster flooding rates to achieve desired convergence goals. This document discusses issues associated with increasing flooding rates and some recommended practices which allow faster flooding rates to be used safely.

Requirements Language

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 January 9, 2022.

Copyright Notice

Copyright (c) 2021 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	2
2. Historical Behavior	3
3. Flooding Rate and Convergence	4
3.1. Flow Control Considerations	5
3.2. Rate of LSP Acknowledgments	7
3.3. Bandwidth Utilization	7
3.4. Packet Prioritization on Receive	7
4. Minimizing LSP Generation	8
5. Redundant Flooding	10
6. Use of Jumbo Frames	10
7. Deployment Considerations	10
8. IANA Considerations	11
9. Security Considerations	11
10. Acknowledgements	11
11. References	11
11.1. Normative References	11
11.2. Informative References	12
Authors' Addresses	12

1. Introduction

Link state IGPs such as Intermediate-System-to-Intermediate-System (IS-IS) depend upon having consistent Link State Databases (LSDB) on all Intermediate Systems (ISs) in the network in order to provide correct forwarding of data packets. When topology changes occur, new/updated Link State PDUs (LSPs) are propagated network-wide. The speed of propagation is a key contributor to convergence time.

Historically, flooding rates have been conservative - on the order of 10s of LSPs/second. This derives from guidance in the base specification [ISO10589] and early deployments when both CPU speeds

and interface speeds were much slower than they are today and the scale of an IS-IS area was smaller than it may be today.

As IS-IS is deployed in greater scale (larger number of nodes in an area and larger number of neighbors/node), the impact of the historic flooding rates becomes more significant. Consider the bringup or failure of a node with 1000 neighbors. This will result in a minimum of 1000 LSP updates. At a typical LSP flooding rate used in many deployments today (33 LSPs/second), it would take 30+ seconds simply to send the updated LSPs to a given neighbor. Depending on the diameter of the network, achieving a consistent LSDB on all nodes in the network could easily take a minute (or more).

Increasing LSP flooding rate therefore becomes an essential element of supporting greater network scale.

The remainder of this document discusses various aspects of protocol operation and how they are impacted by increased flooding rate. Where appropriate, best practices are defined which enhance an implementation's ability to support faster flooding rates.

2. Historical Behavior

The base specification for IS-IS [ISO10589] was first published in 1992 and updated in 2002. The update made no changes in regards to suggested timer values. Convergence targets at the time were on the order of seconds and the specified timer values reflect that. Here are some examples:

minimumLSPGenerationInterval - This is the minimum time interval between generation of Link State PDUs. A source Intermediate system shall wait at least this long before re-generating one of its own Link State PDUs.

The recommended value was 30 seconds.

minimumLSPTransmissionInterval - This is the amount of time an Intermediate system shall wait before further propagating another Link State PDU from the same source system.

The recommended value was 5 seconds.

partialSNPInterval - This is the amount of time between periodic action for transmission of Partial Sequence Number PDUs.

It shall be less than minimumLSPTransmission-Interval.

The recommend value was 2 seconds.

Most relevant to a discussion of LSP flooding rate is the recommended interval between the transmission of two different LSPs on a given interface.

For broadcast interfaces, [ISO10589] defined:

minimumBroadcastLSPTransmissionInterval - the minimum interval between PDU arrivals which can be processed by the slowest Intermediate System on the LAN.

The default value was defined as 33 milliseconds.

NOTE: It was permitted to send multiple LSPs "back-to-back" as a burst, but this was limited to 10 LSPs in a one second period.

Although this value was specific to LAN interfaces, this has commonly been applied by implementations to all interfaces though that was not the original intent of the base specification. In fact Section 12.1.2.4.3 states:

On point-to-point links the peak rate of arrival is limited only by the speed of the data link and the other traffic flowing on that link.

Although modern implementations have not strictly adhered to the 33 millisecond interval, it is commonplace for implementations to limit flooding rate to an order of magnitude similar to the 33 ms value.

In the past 20 years, significant work on achieving faster convergence - more specifically sub-second convergence - has resulted in implementations modifying a number of the above timers in order to support faster signaling of topology changes. For example, minimumLSPGenerationInterval has been modified to support millisecond intervals - often with a backoff algorithm applied to prevent LSP generation storms in the event of a series of rapid oscillations.

However, flooding rate has not been fundamentally altered.

3. Flooding Rate and Convergence

Convergence involves a number of sequential operations.

First the topology change needs to be detected. This is a local activity occurring only on the node or nodes directly connected to the topology change. The directly connected node(s) then must advertise the topology change by updating their LSPs and flooding the changed LSPs. Routers then must process the updated LSDB and

recalculate paths to affected destinations. The updated paths must then be installed in the forwarding plane.

Only when all of the steps are completed on all nodes in the network has the network completed convergence.

As the convergence requirement is consistency of LSDBs on all nodes in the network, it is fundamental to understand that the goal of flooding is to update the LSDB on all nodes in the network "as fast as possible". Controlling the rate of flooding per interface is done to address some practical limitations which include:

- o Fairness to other data and control traffic on the same interface
- o Limitations on the processing rate of incoming control traffic

However, intentionally using different flooding rates on different interfaces increases the possibility of longer periods of LSDB inconsistency, which, in turn, delays network wide convergence.

Many implementations provide knobs to control the rate of LSP flooding on a per interface basis. To the extent that this serves as a flow control mechanism, this may reduce the number of dropped LSPs during high activity bursts and thereby reduce the number of LSP retransmissions required. As LSP retransmission timers are typically long (multiple seconds), this may result in shorter convergence times than if the LSP burst was uncontrolled. But if the performance characteristics of routers in the network are such that some routers consistently accept and process fewer LSPs/second than other routers, convergence will be degraded. Tuning LSP transmission timers on a per interface basis will never provide optimal convergence. Consistent flooding rates should be used on all interfaces.

3.1. Flow Control Considerations

In large scale deployments where an increased flooding rate is being used, it becomes more likely that a burst of LSPs may temporarily overwhelm a receiver. Normal operation of the Update Process will recover from this, but it may well make sense to employ some form of flow control. This will not serve to optimize convergence, but it can serve to reduce the number of LSP retransmissions. As retransmissions are deliberately done at a slow rate, the result of flow control will be to provide a shorter recovery time from a transient condition which prevents a node from handling the targeted rate of LSP transmission. Sustained inability to handle LSP reception at the targeted flooding rate indicates that the network is provisioned in a way which does not support optimal convergence. Steps need to be taken to resolve this issue. Such steps could

include upgrading the routers that demonstrate this condition consistently, altering the configuration on the problematic routers or altering the position of the problematic routers in the network so as to reduce the overall load on those routers, or reducing the target maximum LSP transmission rate network-wide.

When flow control is necessary, it can be implemented in a straightforward manner based on knowledge of the current flooding rate and the current acknowledgement rate. Such an algorithm is a local matter and there is no requirement or intent to standardize an algorithm. There are a number of aspects which serve as guidelines which can be described.

A maximum target LSP transmission rate (LSPTxMax) SHOULD be configurable. This represents the fastest LSP transmission rate which will be attempted. This value SHOULD be applicable to all interfaces and SHOULD be consistent network wide.

When the current rate of LSP transmission (LSPTxRate) exceeds the capabilities of the receiver, the flow control algorithm needs to aggressively reduce the LSPTxRate within a few seconds. Slower responsiveness is likely to result in a large number of retransmissions which can introduce much larger delays in convergence.

NOTE: Even with modest increases in flooding speed (for example, a target LSPTxMax of 300 LSPs/second (10 times the typical rate supported today)), a topology change triggering 2100 new LSPs would only take 7 seconds to complete.

Dynamic adjustment of the rate of LSP transmission (LSPTxRate) upwards (i.e., faster) SHOULD be done less aggressively and only be done when the neighbor has demonstrated its ability to sustain the current LSPTxRate.

The flow control algorithm MUST NOT assume the receive capabilities of a neighbor are static, i.e., it MUST handle transient conditions which result in a slower or faster receive rate on the part of a neighbor.

The flow control algorithm needs to consider the expected delay time in receiving an acknowledgment. See Section 3.2. This may vary per neighbor.

3.2. Rate of LSP Acknowledgments

On point-to-point networks, PSNP PDUs provide acknowledgments for received LSPs. [ISO10589] suggests that some delay be used when sending PSNPs. This provides some optimization as multiple LSPs can be acknowledged in a single PSNP.

If faster LSP flooding is to be used safely, it is necessary that LSPs be acknowledged more promptly as well. This requires a reduction in the delay in sending PSNPs.

As PSNPs also consume link bandwidth and packet queue space and protocol processing time on receipt, the increased sending of PSNPs should be taken into account when considering the rate at which LSPs can be sent on an interface.

3.3. Bandwidth Utilization

Routing protocol traffic has to share bandwidth on a link with other control traffic and data traffic. During periods of instability, routing protocol traffic will increase, but it is still desirable that the maximum bandwidth consumption by routing protocol traffic be modest. This needs to be considered when setting IS-IS flooding rates.

If we assume a maximum size of 1492 bytes for an LSP, here are some rough estimates of bandwidth consumption at different flooding rates:

LSPs/second	100 Mb Link	1 Gb Link
100	1.2 %	0.1 %
500	6.1 %	0.6 %
1000	12.1 %	1.2 %

3.4. Packet Prioritization on Receive

There are three classes of PDUs sent by IS-IS:

- o Hellos
- o LSPs

- o Complete Sequence Number PDUs (CSNPs) and Partial Sequence Number PDUs (PSNPs)

Implementations today may prioritize the reception of Hellos over LSPs and SNPs in order to prevent a burst of LSP updates from triggering an adjacency timeout which in turn would require additional LSPs to be updated.

SNPs serve to acknowledge or trigger the transmission of specified LSPs. On a point-to-point link, PSNPs acknowledge the receipt of one or more LSPs. Because PSNPs (like all IS-IS PDUs) use TLVs in the body, it is possible to acknowledge multiple LSPs using a single PSNP. For this reason, [ISO10589] specifies a delay (partialSNPInterval) before sending a PSNP so that the number of PSNPs required to be sent is reduced. On receipt of a PSNP, the set of LSPs acknowledged by that PSNP can be marked so that they do not need to be retransmitted.

If a PSNP is dropped on reception, this has a significant impact as the set of LSPs advertised in the PSNP cannot be marked as acknowledged and this results in needless retransmissions which may further delay transmission of other LSPs which have yet to be transmitted. It may also make it more likely that a receiver becomes overwhelmed by LSP transmissions.

It is therefore recommended that implementations prioritize the receipt of SNPs over LSPs.

4. Minimizing LSP Generation

In IS-IS the unit of flooding is an LSP. Each router may generate a set of LSPs at each supported level. Each LSP in the set has an LSP number - which is a value from 0-N where N = 255 for the base protocol. (N has been extended to 65535 by [RFC7356].) Each LSP carries network information using defined Type/Length/Value (TLV) tuples. For example, some TLVs carry neighbor information and some TLVs carry reachable prefix information. [ISO10589] strongly recommends preserving the association of a given advertisement (such as a neighbor) with a specific LSP whenever possible. This minimizes the number of LSPs which need to be regenerated when a topology change occurs. This recommendation becomes even more important as the scale of the network increases.

Consider the following example;

Node A has 11 neighbors currently in the UP state and is advertising them in three LSPs with content as follows:

A.00-00 contains the following advertisements

- Neighbor 1
- Neighbor 2
- Neighbor 3
- Neighbor 4
- Neighbor 5

A.00-01 contains the following advertisements:

- Neighbor 6
- Neighbor 7
- Neighbor 8
- Neighbor 9
- Neighbor 10

A.00-02 contains the following advertisements

- Neighbor 11

Imagine that the adjacency to Neighbor 3 goes down. There are (at least) two ways that A could update its LSPs.

Method 1: Node A removes the neighbor advertisement for neighbor 3 from A.00-00 and sends an update for that LSP. LSPs 00-01 and 00-02 are unchanged and so do not have to be flooded.

Method 2: Node A attempts to reduce the number of LSPs currently active and updates the content as follows:

A.00-00 contains the following advertisements

- Neighbor 1
- Neighbor 2
- Neighbor 4
- Neighbor 5
- Neighbor 6

A.00-01 contains the following advertisements:

- Neighbor 7
- Neighbor 8
- Neighbor 9
- Neighbor 10
- Neighbor 11

A.00-02 becomes empty

Node A now has to flood all three LSPs. LSPs #0 and #1 are reflooded because their content has changed. LSP #2 is purged.

In a large scale network, the impact of using Method #2 becomes significant and introduces conditions where a much larger number of LSPs need to be flooded than is the case with Method #1.

In order to operate at scale, implementations need to follow the guidance in [ISO10589] and use Method #1 whenever possible.

5. Redundant Flooding

Default operation of the Update Process is to flood on all interfaces. In cases where a network is highly meshed, this can result in a significant amount of redundant flooding. Nodes will receive multiple copies of each updated LSP.

There are defined mechanisms which can greatly reduce the redundant flooding. These include:

- o Mesh Groups ([RFC2973])
- o Dynamic Flooding ([I-D.ietf-lsr-dynamic-flooding])

6. Use of Jumbo Frames

The maximum size of an LSP (LSPBufferSize) is a parameter that needs to be set consistently network wide. This is because IS-IS does not support fragmentation of its PDUs - so in order for network wide flooding of an LSP to be successful all routers must restrict their LSP size to a size which can be supported without fragmentation on all interfaces on which IS-IS operates.

In networks where all interfaces on which IS-IS operates support large frames, LSPBufferSize may be set to a larger value than the default (1492). This allows more routing information to be encoded in a single LSP, which means that fewer LSPs are generated by each node and therefore the number of LSPs which need to be flooded can be reduced in some scenarios (e.g., node or interface bringup).

7. Deployment Considerations

As noted earlier in this document, it is desired to have consistent flooding speeds on all nodes in the network. Today, this is roughly achieved to the extent that current implementations flood at rates which are on the order of what is discussed in [ISO10589] , i.e., 33 LSPs/second).

As the goal is to introduce an order of magnitude increase in the rate of flooding (e.g., 10 times the current flooding rate) a network which has a mixture of nodes which support the faster flooding speeds and nodes which do not is at greater risk of introducing longer periods of LSDB inconsistency in the network - which is likely to have a negative impact on convergence and increase the occurrence of traffic drops or looping.

It is recommended that all nodes in the network support increased flooding rates before enabling use of the increased flooding rates.

Note that as the Update process runs in the context of an area (or the L2 sub-domain), enablement can safely be done on a per area basis even when nodes in another area do not support the faster flooding rates.

8. IANA Considerations

This document requires no actions by IANA.

9. Security Considerations

Security concerns for IS-IS are addressed in [ISO10589, [RFC5304], and [RFC5310].

10. Acknowledgements

Thanks to Bruno Decraene for his careful review and insightful comments.

11. References

11.1. Normative References

- [ISO10589] International Organization for Standardization, "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, Nov 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>>.
- [RFC2973] Balay, R., Katz, D., and J. Parker, "IS-IS Mesh Groups", RFC 2973, DOI 10.17487/RFC2973, October 2000, <<https://www.rfc-editor.org/info/rfc2973>>.
- [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>>.

- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <<https://www.rfc-editor.org/info/rfc5310>>.
- [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>>.

11.2. Informative References

- [I-D.ietf-lsr-dynamic-flooding]
Li, T., Psenak, P., Ginsberg, L., Chen, H., Przygienda, T., Cooper, D., Jalil, L., Dontula, S., and G. S. Mishra, "Dynamic Flooding on Dense Graphs", draft-ietf-lsr-dynamic-flooding-08 (work in progress), December 2020.
- [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>>.

Authors' Addresses

Les Ginsberg
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
USA

Email: ginsberg@cisco.com

Peter Psenak
Cisco Systems
Apollo Business Center Mlynske nivy 43
Bratislava 821 09
Slovakia

Email: ppsenak@cisco.com

Marek Karasek
Cisco Systems
Pujmanove 1753/10a, Prague 4 - Nusle
Prague 10 14000
Czech Republic

Email: mkarasek@cisco.com

Acee Lindem
Cisco Systems
301 Midenhall Way
Cary, NC 27513
US

Email: acee@cisco.com

Tony Przygienda
Juniper
1137 Innovation Way
Sunnyvale, Ca
USA

Email: prz@juniper.net