

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2022

B. Decraene  
Orange  
C. Bowers  
Jayesh. J  
Juniper Networks, Inc.  
T. Li  
Arista Networks  
G. Van de Velde  
Nokia  
G. Solignac  
Orange  
July 12, 2021

IS-IS Flooding Congestion Control  
draft-decraene-lsr-isis-flooding-speed-07

Abstract

This document proposes a mechanism to adjust IS-IS flooding speed between two adjacent routers by adjusting the sender flooding speed to the capability of the receiver. This helps improving the flooding throughput, reducing LSPs losses and retransmissions due to receiver overload, and avoiding manual tuning of flooding parameters by the network operator. This document defines a new TLV for Hello messages. This TLV may carry a set of parameters indicating the performance capacity to receive LSPs.

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 13, 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 . . . . .	3
1.1. Requirements Language . . . . .	4
2. Overview . . . . .	4
3. Flooding Parameters TLV . . . . .	5
3.1. InterfaceLSPReceiveWindow sub-TLV . . . . .	5
3.2. minimumInterfaceLSPTransmissionInterval sub-TLV . . . . .	5
4. Flow control . . . . .	6
4.1. Operation on a point to point interface . . . . .	6
4.2. Faster acknowledgments of LSPs . . . . .	7
4.3. Operation on a LAN interface . . . . .	8
5. Congestion control . . . . .	9
5.1. Slow start . . . . .	9
5.2. Congestion avoidance . . . . .	10
5.3. Remarks . . . . .	11
6. Interaction with other LSP rate limiting mechanisms . . . . .	11
7. Determining values to be advertised in the Flooding Parameters TLV . . . . .	12
8. Operation considerations . . . . .	13
9. IANA Considerations . . . . .	13
10. Security Considerations . . . . .	13
11. Acknowledgments . . . . .	14
12. References . . . . .	15
12.1. Normative References . . . . .	15
12.2. Informative References . . . . .	15
Appendix A. Changes / Author Notes . . . . .	16
Authors' Addresses . . . . .	17

## 1. Introduction

IGP flooding is paramount for Link State IGP as routing computations assume that the Link State DataBases (LSDBs) are always in sync across all nodes in the flooding domain.

Slow flooding directly translates to delayed network reaction to failure and LSDB inconsistencies across nodes. The former increases packet loss. The latter translates to routing inconsistencies and possibly micro-loops leading to packet loss, link overload, and jitter for all classes of service. Note that across the network, multiple links may be affected by these forwarding issues, even in the case of a single link failure.

In addition, one single event in the network can require the flooding of multiple LSPs. The typical case is a node failure which requires the flooding of at least one LSP per neighbor of the failed node. Hence, if a node has  $N$  IGP neighbors, the failure of this node requires the advertisement and flooding of at least  $N$  LSPs. The network won't be able to converge to the new topology until all  $N$  LSPs are received by all nodes. Hence there is a need to be able to quickly exchange  $N$  LSPs. This document addresses this requirement by allowing the fast flooding of a number of consecutive LSPs.

IGP flooding is hard. One would want fast flooding when the network is stable and slow enough flooding to not overload the neighbor(s) when the network is less stable. Since flooding is performed hop by hop, not overloading the adjacent receiver is sufficient.

Improving the communication speed and efficiency between IS-IS neighbors improves IS-IS scaling. These extensions do not compete with proposed extensions to reduce LSP flooding traffic by reducing the flooding topology such as [I-D.ietf-lsr-dynamic-flooding]. On the contrary, this extension complements those proposals. Indeed reducing the flooding topology does not reduce the size of the LSDB or the total number of LSPs to exchange between two nodes. So increasing the overall flooding speed can be beneficial for nodes implementing dynamic flooding. The reverse is also true: as dynamic flooding reduces the number of neighbors with flooding enabled, this allows nodes implementing the flooding parameter extensions to focus their flooding resources on those neighbors by sending better parameters to the selected flooding nodes and worse parameters to non-selected flooding nodes.

### 1.1. 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 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Overview

Ensuring the goodput between two entities is a layer 4 responsibility as per the OSI model and a typical example is the TCP protocol defined in RFC 793 [RFC0793] . It typically relies on the following sub-functions: flow control, congestion control and reliability.

Flow control is about creating a control loop between a single transmitter and single receiver. TCP provides a mean for the receiver to govern the amount of data sent by the sender. This is achieved by advertising a "receive window", in units of octets, with every ACK. This document proposes to use the same mechanism by advertising a receive window, in units of LSP packets, in IS-IS Hello. The window indicates an allowed number of LSPs that the sender may transmit before receiving acknowledgment of those LSPs. There is an assumption that this is related to the currently available data buffer space available for this adjacency. Indicating a large window encourages transmissions.

Congestion control is about creating multiple interacting control loops between multiple transmitters and multiple receivers. Whereas flow control prevents the sender from overwhelming the receiver, congestion control prevents senders from overwhelming the network. For an IS-IS adjacency, the network between two IS-IS neighbors is relatively limited in scope and consist in a link which is typically over-sized compared to the capability of the IS-IS speakers, but also includes components inside both routers such as a fabric switch, line card CPU and forwarding plane buffers which may experience congestion. This document proposes to use the AIMD (Additive Increase Multiplicative Decrease) algorithm to react to packet loss. Note that TCP Reno relies on the same algorithm.

Reliability relies on loss detection and recovery. IS-IS already has mechanisms to ensure the reliable transmission of LSPs. This is not changed by this document.

### 3. Flooding Parameters TLV

This document defines a new TLV called "Flooding Parameters TLV" that may be included in IIH PDUs. It allows the LSP receiver to advertise receiver related parameters and capabilities which allows the LSP sender to better adapt to the receiver.

Type: TBD1.

Length: variable, the size in octet of the Value field.

Value: a list of sub-TLVs.

Two sub-TLVs are defined in this document.

#### 3.1. InterfaceLSPReceiveWindow sub-TLV

The sub-TLV InterfaceLSPReceiveWindow advertises the maximum number of un-acknowledged LSPs that the node can receive/process with no separation interval between LSPs.

Type: 1.

Length: 4 octets.

Value: number of un-acknowledged LSPs which can be sent back to back.

Note that if an LSP has not been acknowledged and is sent again, it does not count twice. The reason is that this LSP is assumed to be lost and hence not in a buffer at the receiver.

#### 3.2. minimumInterfaceLSPTransmissionInterval sub-TLV

The sub-TLV minimumInterfaceLSPTransmissionInterval advertises the minimum interval, in micro-seconds, between LSPs arrivals which can be processed/received on this interface, after the maximum number of un-acknowledged LSPs has been sent.

Type: 2.

Length: 4 octets.

Value: minimum interval, in micro-seconds, between two consecutive LSPs sent after the receive window has been used.

#### 4. Flow control

Flow control is about creating a control loop between a single transmitter and single receiver. This document proposes to use a mechanism similar to the TCP receive window to allow the receiver to govern the amount of data sent by the sender. This receive window indicates an allowed number of LSPs that the sender may transmit before receiving acknowledgment of those LSPs. This receive window, in units of LSPs, is advertised in the sub-TLV InterfaceLSPReceiveWindow.

##### 4.1. Operation on a point to point interface

By sending the InterfaceLSPReceiveWindow sub-TLV with a value N, the node advertises to its IS-IS neighbor, its ability to receive a maximum of N un-acknowledged LSPs from this neighbor, with no separation interval. This is akin to a reception window or sliding window in flow control. This value typically reflects the socket buffer size. Special care must be taken to let space for Hello and SNP PDUs if they share the same socket. In this case, this document suggests to advertise a Receive Window corresponding to half the size of the socket buffer.

By sending the minimumInterfaceLSPTransmissionInterval sub-TLV with a value T, the node advertises to its IS-IS neighbor, its ability to receive, after the receive window is full, LSPs separated by at least T micro-seconds from this neighbor.

The IS transmitter MUST NOT exceed these parameters. After having sent N un-acknowledged LSPs, it MUST send the following LSPs with an interval of at least T micro-seconds between each LSP.

Note however that if either the LSP transmitter or receiver does not adhere to these parameters, for example because of transient conditions, this causes no fatal condition to the operation of IS-IS. The worst case, the loss of LSP on the IS receiver, is already accounted for in [ISO10589]. As per [ISO10589], after a few seconds, respectively 2 and 10 by default in [ISO10589], neighbors will exchange PSNP (for point to point interface) or CSNP (for broadcast interface) and recover from the lost LSPs. This worst case (overrunning the receiver) should however be avoided as those additional seconds are impacting the network and the traffic as the LSDB is not fully synchronized. Hence it is better to err on the conservative side and to under-run the receiver rather than over-run it.

For a given IS-IS adjacency, the Flooding Parameters TLV does not need to be advertised in each IIH. The IS transmitter uses the

latest received value for each parameter (sub-TLV) until a new value is advertised by the IS receiver. Note however that IIH are not reliably exchanged, hence may never be received. For a parameter which has never been advertised, the IS transmitter use its local default value. That value SHOULD be configurable on a per node basis and MAY be configurable on a per interface basis.

#### 4.2. Faster acknowledgments of LSPs

As per [ISO10589] , on point to point interfaces, the LSP receiver dynamically acknowledges the received LSPs by sending PSNP messages.

By acknowledging the LSPs before the InterfaceLSPReceiveWindow is exhausted, the receiver can achieve dynamic flow control and increase the flooding throughput without risking overloading any IS-IS router. If the InterfaceLSPReceiveWindow is large enough, the downstream flooding node can acknowledge a set of multiple LSPs up to the maximum size of a PSNP (90 LSPs) which allows dynamic flow control with limited or even no increase in the number of sent PSNPs.

In order to avoid reducing the throughput, the receiver should avoid letting the receive window exhaust. Therefore, the receiver SHOULD acknowledge the LSP more quickly than the default specified in [ISO10589] . This is beneficial both to the LSP sender which receives faster feedback and to the LSP receiver which has more time to acknowledge many LSPs before the sender times out and resend the LSP.

Receiver SHOULD reduce partialSNPInterval. The choice of this lower value is a local choice. It may depend on the (available) processing power of the node, the number of adjacencies, the requirement to synchronize the LSDB more quickly. 200 ms seems a reasonable value.

In addition to the timer based partialSNPInterval, the receiver SHOULD keep track of the number of unacknowledged LSPs per circuit and level. When this number exceeds a preset threshold LSP per PSNP (LPP), the receiver SHOULD immediately send a PSNP without waiting for the PSNP timer to expire. In case of a burst of LSPs, this allows for more frequent PSNPs, hence a faster feedback loop to the sender. In the absence of burst, the usual time-based PSNP approach comes into effect. This number SHOULD be lower than the advertised receive window InterfaceLSPReceiveWindow, e.g., InterfaceLSPReceiveWindow/2. This number SHOULD also be lower or equal to 90 as this is the maximum number of LSPs that can be acknowledged in a PSNP, hence waiting longer would not reduce the number of PSNPs sent but would delay the acknowledgements. Best performance is achieved when this number is an integer fraction of InterfaceLSPReceiveWindow. Based on experimental evidence, 15

unacknowledged LSPs is a right value assuming that InterfaceLSPReceiveWindow is at least twice bigger (>30).

By deploying both the time-based and the threshold-based PSNP approaches, the receiver can be adaptive to both LSP bursts and infrequent LSP updates.

#### 4.3. Operation on a LAN interface

On a LAN interface, an IS receiver will generally receive LSPs from multiple IS transmitters. Also the LSPs sent by a given IS transmitter is received by all of the IS receivers on that LAN. In this section, we clarify how the flooding parameters should be interpreted in the context of a LAN.

An IS receiver on a LAN will communicate its desired flooding parameters using a single Flooding Parameters TLV, copies of which will be received by all N transmitters. The flooding parameters sent by the IS receiver MUST be understood as instructions from the receiver to each transmitter about the desired maximum transmit characteristics of each transmitter. The receiver is aware that there are N transmitters that can send LSPs to the receiver LAN interface. The receiver might want to take that into account by advertising a higher value of InterfaceLSPTransmissionInterval on this LAN interface than what it would advertise on a point to point interface. When the transmitters receive the InterfaceLSPTransmissionInterval value advertised by the DIS receiver, the transmitters should rate limit LSPs according to the advertised flooding parameters. They should not apply any further interpretation to the flooding parameters advertised by the receiver.

A given IS transmitter will receive flooding parameter advertisements from N different Flooding Parameters TLVs, which may carry different flooding parameter values. A given transmitter SHOULD use the most conservative value on a per Flooding parameter basis. For example, if the transmitter receives InterfaceLSPReceiveWindow from N IS-IS nodes on the LAN, it should use the smallest value.

In order for the InterfaceLSPReceiveWindow to be a useful parameter, an IS transmitter needs to be able to keep track of the number of unacknowledged LSPs it has sent to a given IS receiver. On a LAN there is no explicit acknowledgment of the receipt of LSPs between a given IS transmitter and a given IS receiver. However, an IS transmitter on a LAN can infer whether any IS receiver on the LAN has requested retransmission of LSPs from the DIS, by monitoring PSNPs generated on the LAN. If no PSNPs have been generated on the LAN for a suitable period of time, then an IS transmitter can safely set the number of unacknowledged LSPs to zero. Since this suitable period of time is



much higher than the fast acknowledgment of LSP defined in Section 4.2 , the sustainable sending rate of LSP will be much slower on a LAN interface compared to a point to point interface. However, InterfaceLSPReceiveWindow is still very useful for the first LSPs sent and hence usefull for the faster flooding in case of a single node failure which requires to flood a relatively small number of LSPs.

A compliant implementation may choose to not support this operation on a LAN interface.

## 5. Congestion control

Whereas flow control prevents the sender from overwhelming the receiver, congestion control prevents senders from overwhelming the network. For an IS-IS adjacency, the network between two IS-IS neighbors is relatively limited in scope and includes a single link which is typically over-sized compared to the capability of the IS-IS speakers. It also includes components inside both routers such as a fabric switch, line cards CPU and forwarding plane buffers which may experience congestion. This document proposes one optional congestion control algorithm but implementations may choose a different one or none.

The congestion control algorithm defined in this document is largely inspired by the TCP congestion control algorithm RFC 5681 [RFC5681]. A congestion control algorithm is comprised of three elements : a slow start phase, a congestion avoidance phase, and a transition between the two.

The proposed algorithm uses a variable Congestion window 'cwin'. It plays the same role as Receive Window described before. The main difference is that CWin is dynamically changed according to the feedback obtained by the PSNPs.

### 5.1. Slow start

The goal of the slow start phase is to grow fast and try to estimate the effective link capacity.

The algorithm is circuit scoped. At the beginning of the slow start, the sender starts with:

- o a congestion window (cwin) set to one. `cwin := 1;`
- o a number of acked LSPs. `acked_lsps := 0;`
- o a max seen bandwidth. `max_bw := 0;`

o a current rtt estimate. `cur_rtt := NA;`

Upon LSP sending, a sender records for said LSP the current time in `time_sent` and `acked_lsps` in `acked_lsps_sent`. This data is tied to each LSP.

Upon PSNP reception, a sender does the following:

```

cwin := min(cwin + nb_of_lsp_entries, rwin)
acked_lsps += nb_of_lsp_entries
max_diff := 0
max_bw := 0
for every LSP entry:
    time_to_ack := time_now - time_sent
    nb_acked := acked_lsps - acked_lsps_sent
    bw_est := nb_acked/time_to_ack
    max_bw := max(max_bw, bw_est)
    max_diff := max(max_diff, time_to_ack)

if cur_rtt == NA then cur_rtt = max_diff
else cur_rtt := 7/8 * cur_rtt + 1/8 * max_diff

```

Figure 1

Starting with the first PSNP, `max_bw` is checked every `cur_rtt`. Once it has stalled for 3 consecutive times, the congestion control algorithm transitions from slow start to congestion avoidance. There is bandwidth stalling when the bandwidth has not increased by at least 25% compared the last RTT. Note that this is similar to Google's BBR ([I-D.cardwell-iccr-g-bbr-congestion-control] ) slow start phase.

## 5.2. Congestion avoidance

The goal of the congestion avoidance phase is to try to stay close to the effective capacity of the link. For this, the algorithm estimates the maximum time taken by the receiver to acknowledge a LSP. If an LSP arrives slower than this delay, congestion is inferred and `cwin` is decreased.

Upon PSNP reception, a sender does the following:

```
cwin = min(cwin + N/congestion window, rwin)
rtt_est := 0
for every LSP entry:
    time_to_ack = time_now - time_sent
    rtt_est = max(rtt_est, time_to_ack)

if rtt_var == NA then rtt_var = rtt_est / 2
else rtt_var = 3/4 * rtt_var + 1/4 * abs(cur_rtt - rtt_est)

cur_rtt = 7/8 * cur_rtt + 1/8 * rtt_est
```

Figure 2

Every LSP is checked to be acked within  $\text{cur\_rtt} + \text{rtt\_var}$ . If an LSP arrives late,  $\text{cwin}$  is divided by two. This behaviour is similar to TCP retransmission timer defined in RFC 6298 [RFC6298]

Note: there is no need for a timer per LSP. A timer per RTT is enough. During an RTT, sent LSPs are recorded in a list `list_1`. Once the RTT is over, `list_1` is kept and another list `list_2` is used to store the next LSPs. LSPs are removed from the lists when acked. At the end of the second RTT, every LSP in `list_1` should have been acked, so `list_1` is checked to be empty. `list_1` can then be reused for the next RTT.

If there is no transmitted LSP for a fixed period of time, e.g. 2 seconds, the sender switches back to the slow start phase.

### 5.3. Remarks

This algorithm's performance is dependent on the LPP value. Indeed, the smaller LPP is, the more information is available for the congestion control algorithm to perform well. However, it also increases the resources spent on sending PSNPs, so a tradeoff must be made. This document recommends to use an LPP of 15 or less.

Note that this congestion control algorithm benefits from the extensions proposed in this document. The advertisement of a receive window from the receiver ( Section 4 ) avoids the use of an arbitrary maximum value by the sender. The faster acknowledgment of LSP ( Section 4.2 ) allows for a faster control loop and hence a faster increase of the congestion window in the absence of congestion.

### 6. Interaction with other LSP rate limiting mechanisms

[ISO10589] describes a mechanism that limits the rate at which LSPs from the same source system are sent out on interfaces. (See the description of the parameter

minimumBroadcastLSPTranLSPTransmissionInterval in section 7.3.15.6 of [ISO10589] ). In practice, however, router vendors have implemented mechanisms that limit the rate of LSPs sent on a given interface. This is often configurable on a per-interface basis using 'lsp-interval' or 'lsp-pacing-interval' CLI configuration). The mechanism described in the current document extends the practice of limiting the rate of LSPs sent on a given interface, by using parameters advertised by the LSP receiver. When the mechanism described in the current document is used, the mechanism described in section 7.3.15.6 of [ISO10589] is not used.

#### 7. Determining values to be advertised in the Flooding Parameters TLV

The values that a receiving IS advertises do not need to be close to perfection. It is OK to be too low and hence not to use the full bandwidth or CPU resources. It is OK to be too high during some situation and hence have the receiver drop some LSPs as the IS-IS protocol has mechanisms to recover. What is not OK is to flood multiple order of magnitudes slower than both nodes can achieve, or to consistently overload the receiver.

The values may not need to be dynamic as a form of dynamic is provided by the dynamic acknowledgment of LSPs in SNP messages. Acknowledgments provides a feedback loop on how fast/slower the LSPs are processed by the receiver. They also signal that the LSPs have been processed by the receiver hence removed from receive window, explicitly signaling to the sender that more LSPs may be sent. By advertising relatively static parameters, we expect to produce overall flooding behavior similar to what might be achieved by manually configuring per-interface LSP rate limiting on all interfaces in the network. The advertised values may be based, for example, on an off line tests of the overall LSP processing speed for a particular set of hardware and the number of interfaces configured for IS-IS. With such a formula, the values advertised in the Flooding Parameters TLV would only change when additional IS-IS interfaces are configured.

The values may be updated dynamically, to reflect the relative change of load of the receiver, by improving the values when the receiver load is getting lower and degrading the values when the receiver load is getting higher. For example, if LSPs are regularly dropped, or the queue regularly comes close to being filled, then values may be too high. On the other hand, if the queue is barely used (by IS-IS), then values may be too low.

The values may also be absolute value reflecting relevant (averaged) hardware resources that are been monitored, typically the amount of buffer space used by incoming LSPs. In this case, care must be taken

when choosing the parameters influencing the values, in order to avoid undesirable or instable feedback loops. It would be undesirable to use a formula that depends, for example, on an active measurement of the instantaneous CPU load to modify the values advertised in the Flooding Parameters TLV. This could introduce feedback into the IGP flooding process that could produce unexpected behavior.

## 8. Operation considerations

As discussed in Section 4.3 , the solution is more effective on point to point adjacencies. Hence a broadcast interface (e.g. Ethernet) only shared by two IS-IS neighbors should be configured as point to point in order to have a more effective flooding.

## 9. IANA Considerations

IANA is requested to allocate one TLV from the IS-IS TLV codepoint registry.

Type	Description	IIH	LSP	SNP	Purge
----	-----	---	---	---	---
TBD1	Flooding Parameters TLV	y	n	n	n

Figure 3

This document creates the following sub-TLV Registry:

Name: Sub-TLVs for TLV TBD1 (Flooding Parameters TLV).

Registration Procedure: Expert Review [RFC8126] .

Type	Description
0	Reserved
1	InterfaceLSPReceiveWindow
2	minimumInterfaceLSPTransmissionInterval
3-255	Unassigned

Table 1: Initial allocations

## 10. Security Considerations

Any new security issues raised by the procedures in this document depend upon the ability of an attacker to inject a false but

apparently valid SNP or IIH, the ease/difficulty of which has not been altered.

As with others TLV advertisements, the use of a cryptographic authentication as defined in [RFC5304] or [RFC5310] allows the authentication of the peer and the integrity of the message. As this document defines a TLV for SNP or IIH message, the relevant cryptographic authentication is for SNP and IIH message.

In the absence of cryptographic authentication, as IS-IS does not run over IP but directly over the link layer, it's considered difficult to inject false SNP/IIH without having access to the link layer.

If a false SNP/IIH is sent with a Flooding Parameters TLV set to conservative values, the attacker can reduce the flooding speed between the two adjacent neighbors which can result in LSDB inconsistencies and transient forwarding loops. However, it is not significantly different than filtering or altering LSPDUs which would also be possible with access to the link layer. In addition, if the downstream flooding neighbor has multiple IGP neighbors, which is typically the case for reliability or topological reasons, it would receive LSPs at a regular speed from its other neighbors and hence would maintain LSDB consistency.

If a false SNP/IIH is sent with a Flooding Parameters TLV set to aggressive values, the attacker can increase the flooding speed which can either overload a node or more likely generate loss of LSPs. However, it is not significantly different than sending many LSPs which would also be possible with access to the link layer, even with cryptographic authentication enabled. In addition, IS-IS has procedures to detect the loss of LSPs and recover.

This TLV advertisement is not flooded across the network but only sent between adjacent IS-IS neighbors. This would limit the consequences in case of forged messages, and also limits the dissemination of such information.

## 11. Acknowledgments

The authors would like to thank Henk Smit, Sarah Chen, Xuesong Geng, Pierre Francois and Hannes Gredler for their reviews, comments and suggestions.

The authors would like to thank David Jacquet, Sarah Chen, and Qiangzhou Gao for the tests performed on commercial implementations and their identification of some limiting factors.

## 12. References

### 12.1. Normative References

- [ISO10589]  
International Organization for Standardization,  
"Intermediate system to Intermediate system intra-domain  
routing 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>>.
- [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>>.
- [RFC6298] Paxson, V., Allman, M., Chu, J., and M. Sargent,  
"Computing TCP's Retransmission Timer", RFC 6298,  
DOI 10.17487/RFC6298, June 2011,  
<<https://www.rfc-editor.org/info/rfc6298>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for  
Writing an IANA Considerations Section in RFCs", BCP 26,  
RFC 8126, DOI 10.17487/RFC8126, June 2017,  
<<https://www.rfc-editor.org/info/rfc8126>>.
- [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>>.

### 12.2. Informative References

- [I-D.cardwell-iccr-g-bbr-congestion-control]  
Cardwell, N., Cheng, Y., Yeganeh, S. H., and V. Jacobson,  
"BBR Congestion Control", draft-cardwell-iccr-g-bbr-  
congestion-control-00 (work in progress), July 2017.

[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.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<https://www.rfc-editor.org/info/rfc793>>.

[RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, DOI 10.17487/RFC5681, September 2009, <<https://www.rfc-editor.org/info/rfc5681>>.

#### Appendix A. Changes / Author Notes

[RFC Editor: Please remove this section before publication]

00: Initial version.

01: Two notes added in section 3 "Operation".

02: Refresh, no technical change.

03:

- o Flooding Parameters TLV: name changed, advertised in both Hello and SNP rather than just Hello, contains sub-TLVs, parameters encoded in 4 octets.
- o Terminology: upstream/downstream terms removed, in favor of terms from ISO specification (transmitter, receiver); burst-size rename to receive-window.
- o Significant editorials changes.
- o New section on the faster acknowledgment of LSPs.
- o New section on the faster retransmission of lost LSPs.

04:

- o Adding general introduction on flow control, congestion control, loss detection and recovery.
- o Reorganizing sections as per the high level functions: flow control, congestion control, loss detection and recovery.



- o Adding a section on congestion control.

05:

- o Some editorials changes.
- o Updating section "Faster acknowledgments of LSPs" following the IS-IS flooding performance tests presented during IETF 108.
- o Updated IANA section (new registry).

06: Refresh, no technical change.

07:

- o Precision that if a LSP is lost and resent, it does not count twice in the InterfaceLSPReceiveWindow.
- o Title changed.
- o Removed fast retransmissions of LSPs.
- o Changed congestion control algorithm.
- o Removed support of TLV in SNP.

#### Authors' Addresses

Bruno Decraene  
Orange

Email: [bruno.decraene@orange.com](mailto:bruno.decraene@orange.com)

Chris Bowers  
Juniper Networks, Inc.  
1194 N. Mathilda Avenue  
Sunnyvale, CA 94089  
USA

Email: [cbowers@juniper.net](mailto:cbowers@juniper.net)

Jayesh J  
Juniper Networks, Inc.  
1194 N. Mathilda Avenue  
Sunnyvale, CA 94089  
USA

Email: jayeshj@juniper.net

Tony Li  
Arista Networks  
5453 Great America Parkway  
Santa Clara, California 95054  
USA

Email: tony.li@tony.li

Gunter Van de Velde  
Nokia  
Copernicuslaan 50  
Antwerp 2018  
Belgium

Email: gunter.van\_de\_velde@nokia.com

Guillaume Solignac  
Orange

Email: guillaume.solignac@orange.com

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](mailto:ginsberg@cisco.com)

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

Email: [ppsenak@cisco.com](mailto: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

SPRING  
Internet-Draft  
Intended status: Standards Track  
Expires: 24 September 2022

S. Hegde  
W. Britto  
R. Shetty  
Juniper Networks Inc.  
B. Decraene  
Orange  
P. Psenak  
Cisco Systems  
T. Li  
Arista Networks  
23 March 2022

Flexible Algorithms: Bandwidth, Delay, Metrics and Constraints  
draft-ietf-lsr-flex-algo-bw-con-02

Abstract

Many networks configure the link metric relative to the link capacity. High bandwidth traffic gets routed as per the link capacity. Flexible algorithms provides mechanisms to create constraint based paths in IGP. This draft documents a generic metric type and set of bandwidth related constraints to be used in Flexible Algorithms.

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 [RFC2119].

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 24 September 2022.



## Copyright Notice

Copyright (c) 2022 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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Generic Metric Advertisement . . . . .	4
2.1. ISIS Generic Metric sub-TLV . . . . .	5
2.2. OSPF Generic Metric sub-TLV . . . . .	6
2.3. Generic Metric applicability to Flexible Algorithms Multi-domain/Multi-area networks . . . . .	7
3. FAD constraint sub-TLVs . . . . .	7
3.1. ISIS FAD constraint sub-TLVs . . . . .	8
3.1.1. ISIS Exclude Minimum Bandwidth sub-TLV . . . . .	8
3.1.2. ISIS Exclude Maximum Delay sub-TLV . . . . .	9
3.2. OSPF FAD constraint sub-TLVs . . . . .	10
3.2.1. OSPF Exclude Minimum Bandwidth sub-TLV . . . . .	10
3.2.2. OSPF Exclude Maximum Delay sub-TLV . . . . .	11
4. Bandwidth Metric Advertisement . . . . .	11
4.1. Automatic Metric Calculation . . . . .	12
4.1.1. Automatic Metric Calculation Modes . . . . .	12
4.1.2. Automatic Metric Calculation Methods . . . . .	13
4.1.3. ISIS FAD constraint sub-TLVs for automatic metric calculation . . . . .	14
4.1.4. OSPF FAD constraint sub-TLVs for automatic metric calculation . . . . .	18
5. Bandwidth metric considerations . . . . .	22
6. Calculation of Flex-Algorithm paths . . . . .	23
7. Backward Compatibility . . . . .	23
8. Security Considerations . . . . .	23
9. IANA Considerations . . . . .	23
9.1. IGP Metric-Type Registry . . . . .	23
9.2. ISIS Sub-Sub-TLVs for Flexible Algorithm Definition Sub-TLV . . . . .	23
9.3. OSPF Sub-TLVs for Flexible Algorithm Definition Sub-TLV . . . . .	24
9.4. Sub-TLVs for TLVs 22, 23, 25, 141, 222, and 223 . . . . .	24

9.5. OSPFv2 Extended Link TLV Sub-TLVs . . . . .	25
9.6. Types for sub-TLVs of TE Link TLV (Value 2) . . . . .	25
9.7. OSPFv3 Extended-LSA Sub-TLVs . . . . .	25
10. Acknowledgements . . . . .	25
11. Contributors . . . . .	25
12. References . . . . .	25
12.1. Normative References . . . . .	25
12.2. Informative References . . . . .	26
Authors' Addresses . . . . .	27

## 1. Introduction

High bandwidth traffic such as residential internet traffic and machine to machine elephant flows benefit from using high capacity links. Accordingly, many network operators define a link's metric relative to its capacity to help direct traffic to higher bandwidth links, but this is no guarantee that lower bandwidth links will be avoided, especially in failure scenarios. To ensure that elephant flows are only placed on high capacity links, it would be useful to explicitly exclude the high bandwidth traffic from utilizing links below a certain capacity. Flex-Algorithm [I-D.ietf-lsr-flex-algo] has already defined as a set of parameters consisting of calculation-type, metric-type and a set of constraints for allowing operators to have more control over network path computation. In this document, we define further extensions to Flex-Algorithm that will allow operators additional control over their traffic, especially with respect to constraints about bandwidth.

Historically, IGPs have done path computation by minimizing the sum of the link metrics along the path from source to destination. While the metric has been administratively defined, implementations have defaulted to a metric that is inversely proportional to link bandwidth. This has driven traffic to higher bandwidth links and has required manual metric manipulation to achieve the desired loading of the network.

Over time, with the addition of different traffic types, the need for alternate types of metrics has become clear. Flex-Algorithm already supports using the minimum link delay and the administratively assigned traffic-engineering metrics in path computation. However, it is clear that additional metrics may be of interest in different situations. A network operator may seek to minimize their operational costs and thus may want a metric that reflects the actual fiscal costs of using a link. Other traffic may require low jitter, leading to an entirely different set of metrics. With Flex-Algorithm, all of these different metrics, and more, could be used concurrently on the same network.

In some circumstances, path computation constraints, such as administrative groups, can be used to ensure that traffic avoids particular portions of the network. These strict constraints are appropriate when there is an absolute requirement to avoid parts of the topology, even in failure conditions. If, however, the requirement is less strict, then using a high metric in a portion of the topology may be more appropriate.

This document defines a family of generic metrics that can carry various types of administratively assigned metrics. This document proposes standard metric-types which require specific standard document. This document also proposes user defined metric-types where specifics are not defined, so that administrators are free to assign semantics as they fit. This document also specifies a new bandwidth based metric type to be used with Flex-Algorithm and other applications in Section Section 4. Additional Flexible Algorithm Definition (FAD) constraints are defined in Section Section 3 that allow the network administrator to preclude the use of low bandwidth links or high delay links. Section Section 4.1 defines mechanisms to automatically calculate link metrics based on parameters defined in the FAD and the advertised Maximum Link Bandwidth of each link. This is advantageous because administrators can change their criteria for metric assignment centrally, without individual modification of each link metric throughout the network.

## 2. Generic Metric Advertisement

ISIS and OSPF advertise a metric for each link in their respective link state advertisements. Multiple metric types are already supported. Administratively assigned metrics are described in the original OSPF and ISIS specifications. The Traffic Engineering Default Metric is defined in [RFC5305] and [RFC3630] and the Min Unidirectional delay metric is defined in [RFC8570] and [RFC7471]. Other metrics, such as jitter, reliability, and fiscal cost may be helpful, depending on the traffic class. Rather than attempt to enumerate all possible metrics of interest, this document specifies a generic mechanism for advertising metrics.

Each generic metric advertisement is on a per-link and per metric type basis. The metric advertisement consists of a metric type field and a value for the metric. The metric type field is assigned by the "IGP metric type" IANA registry. Metric types 0-127 are standard metric types as assigned by IANA. This document further specifies a user defined metric type space of metric types 128-255. These are user defined and can be assigned by an operator for local use.

## 2.1. ISIS Generic Metric sub-TLV

The ISIS Generic Metric sub-TLV specifies the link metric for a given metric type. Typically, this metric is assigned by a network administrator. Generic metric is application-independent attribute similar to igp-metric. The Generic Metric sub-TLV is advertised in the TLVs/sub-TLVs below:

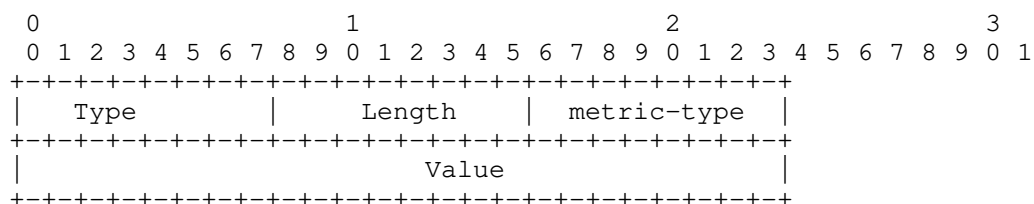
TLV-22 (Extended IS reachability) [RFC5305]

TLV-222 (MT-ISN) [RFC5120]

TLV-23 (IS Neighbor Attribute) [RFC5311]

TLV-223 (MT IS Neighbor Attribute) [RFC5311]

TLV-141 (inter-AS reachability information) [RFC5316]



Type : TBD (To be assigned by IANA)

Length: 4 octets

metric-type: A value from the IGP metric-type registry

Value : metric value range (1 - 16,777,215)

Figure 1: ISIS Generic Metric sub-TLV

The Generic Metric sub-TLV MAY be advertised multiple times. For a particular metric type, the Generic Metric sub-TLV MUST be advertised only once for a link when advertised in TLV 22,222,23,223 and 141. If there are multiple Generic Metric sub-TLVs advertised for a link for same metric type in one or more received LSPDUs, advertisement in the lowest numbered fragment MUST be used and the subsequent ones MUST be ignored. If the metric type indicates a standard metric type for which there are other advertisement mechanisms (e.g., the IGP metric, the Min Unidirectional Link Delay, or the Traffic Engineering Default Metric, as of this writing), the Generic Metric advertisement MUST be ignored.

## 2.2. OSPF Generic Metric sub-TLV

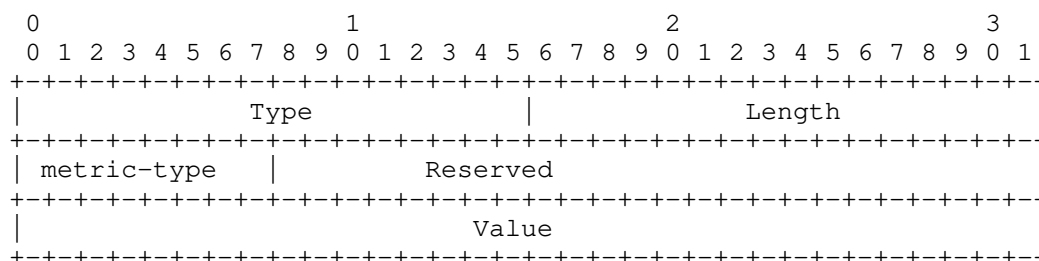
The OSPF Generic Metric sub-TLV specifies the link metric for a given metric type. Typically, this metric is assigned by a network administrator. Generic metric is application-independent attribute similar to igp-metric. The Generic Metric sub-TLV is advertised in the TLVs below:

sub-TLV of the OSPF Link TLV of OSPF extended Link LSA [RFC7684].

sub-TLV of TE Link TLV (2) of OSPF TE LSA [RFC3630].

sub-TLV of the Router-Link TLV in the E-Router-LSA in OSPFv3 [RFC8362].

The Generic Metric sub-TLV is TLV type TBD (IANA), and is eight octets in length.



Type : TBD (To be assigned by IANA)

Length: 8 octets

metric-type = A value from the IGP metric type registry

Value : metric value (1- 4,294,967,295)

Figure 2: OSPF Generic Metric sub-TLV

The Generic Metric sub-TLV MAY be advertised multiple times. For a particular metric type, the Generic Metric sub-TLV MUST be advertised only once for a link when advertised in OSPF Link TLV of Extended Link LSA, Link TLV of TE LSA and sub-TLV of the Router-Link TLV in the E-Router-LSA Router-Link TLV in OSPFv3. If there are multiple Generic Metric sub-TLVs advertised for a link for the same metric type in a received LSA, the first one MUST be used and the subsequent ones MUST be ignored. If the metric type indicates a standard metric type for which there are other advertisement mechanisms (e.g., the IGP metric, the Min Unidirectional Link Delay, or the Traffic Engineering Default Metric, as of this writing), the Generic Metric advertisement MUST be ignored.

### 2.3. Generic Metric applicability to Flexible Algorithms Multi-domain/Multi-area networks

Generic Metric can be used by Flex-Algorithms by specifying the metric type in the Flexible Algorithm Definitions. When Flex-Algorithms is used in a multi-area network, [I-D.ietf-lsr-flex-algo] defines FAPM sub-TLV that carries the Flexible Algorithm specific metric. Metric carried in FAPM will be equal to the metric to reach the prefix for that Flex-Algorithm in its source area or domain. When Flex-Algorithm uses Generic metric, the same procedures as described in section 13 of [I-D.ietf-lsr-flex-algo] are used to send and process FAPM sub-TLV.

### 3. FAD constraint sub-TLVs

In networks that carry elephant flows, directing an elephant flow down a low-bandwidth link would be catastrophic. Thus, in the context of Flex-Algorithm, it would be useful to be able to constrain the topology to only those links capable of supporting a minimum amount of bandwidth.

If the capacity of a link is constant, this can already be achieved through the use of administrative groups. However, when a Layer 3 link is actually a collection of Layer 2 links (LAG/Layer 2 Bundle), the link bandwidth will vary based on the set of active constituent links. This could be automated by having an implementation vary the advertised administrative groups based on bandwidth, but this seems unnecessarily complex and expressing this requirement as a direct constraint on the topology seems simpler. This is also advantageous if the minimum required bandwidth changes, as this constraint would provide a single centralized, coordinated point of control.

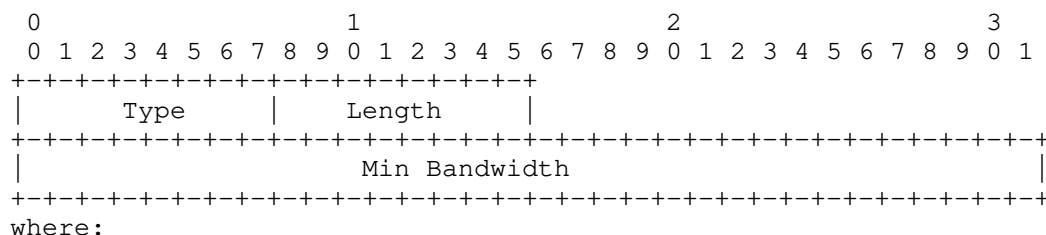
To implement this idea, this document defines a new Exclude Minimum Bandwidth constraint. When this constraint is advertised in a FAD, a link will be pruned from the Flex-Algorithm topology if the link's advertised Maximum Link Bandwidth is below the advertised Minimum Bandwidth value.

Similarly, this document defines a Exclude Maximum Link Delay constraint. Delay is an important consideration in High Frequency Trading applications, networks with transparent L2 link recovery, or in satellite networks, where link delay may fluctuate. Mechanisms already exist to measure the link delay dynamically and advertised it in the IGP. Networks that employ dynamic link delay measurement, may want to exclude links that have a delay over a given threshold.

### 3.1. ISIS FAD constraint sub-TLVs

#### 3.1.1. ISIS Exclude Minimum Bandwidth sub-TLV

ISIS Flex-Algorithm Exclude Minimum Bandwidth sub-TLV (FAEMB) is a sub-TLV of the ISIS FAD sub-TLV. It has the following format:



Type: TBA

Length: 4 octets.

Min Bandwidth: The link bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.

Figure 3: ISIS FAEMB sub-TLV

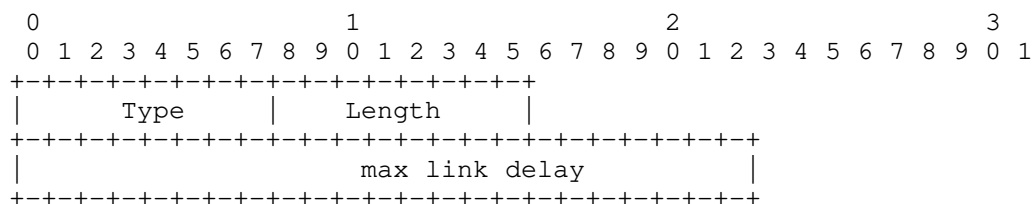
The FAEMB sub-TLV MUST appear at most once in the FAD sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

The Minimum bandwidth advertised in FAEMB sub-TLV MUST be compared with Maximum Link Bandwidth advertised in sub-sub-TLV 9 of ASLA sub-TLV [RFC 8919]. If L-Flag is set in the ASLA sub-TLV, the Minimum bandwidth advertised in FAEMB sub-TLV MUST be compared with Maximum Link Bandwidth as advertised by the sub-TLV 9 of the TLV 22/222/23/223/141 [RFC 5305] as defined in [RFC8919] Section 4.2.

If the Maximum Link Bandwidth is lower than the Minimum link bandwidth advertised in FAEMB sub-TLV, the link MUST be excluded from the Flex-Algorithm topology. If a link does not have the Maximum Link Bandwidth advertised but the FAD contains this sub-TLV, then that link MUST NOT be excluded from the topology based on the Minimum Bandwidth constraint.

### 3.1.2. ISIS Exclude Maximum Delay sub-TLV

ISIS Flex-Algorithm Exclude Maximum Delay sub-TLV (FAEMD) is a sub-TLV of the ISIS FAD sub-TLV. It has the following format.



where:

Type: TBD

Length: 3 octets

Max link delay: Maximum link delay in microseconds

Figure 4: ISIS FAEMD sub-TLV

The FAEMD sub-TLV MUST appear only once in the FAD sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

The Maximum link delay advertised in FAEMD sub-TLV MUST be compared with Min Unidirectional Link Delay advertised in sub-sub-TLV 34 of ASLA sub-TLV [RFC 8919]. If L-Flag is set in the ASLA sub-TLV, the Maximum link delay advertised in FAEMD sub-TLV MUST be compared with Min Unidirectional Link Delay as advertised by the sub-TLV 34 of the TLV 22/222/23/223/141 [RFC 8570] as defined in [RFC8919] Section 4.2.

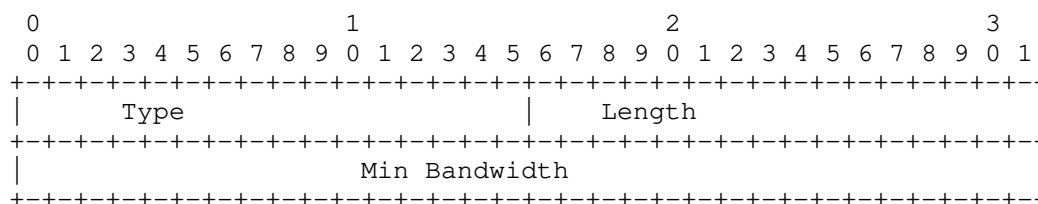


If the Min Unidirectional Link Delay value is higher than the Maximum link delay advertised in FAEMD sub-TLV, the link MUST be excluded from the Flex-Algorithm topology. If a link does not have the Min Unidirectional Link Delay advertised but the FAD contains this sub-TLV, then that link MUST NOT be excluded from the topology based on the Maximum Delay constraint.

### 3.2. OSPF FAD constraint sub-TLVs

#### 3.2.1. OSPF Exclude Minimum Bandwidth sub-TLV

OSPF Flex-Algorithm Exclude Minimum Bandwidth sub-TLV (FAEMB) is a sub-TLV of the OSPF FAD TLV. It has the following format.



where:

Type: TBD

Length: 4 octets.

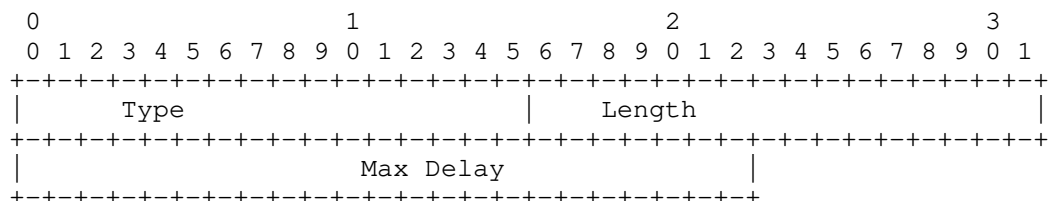
Min Bandwidth: link bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.

Figure 5: OSPF FAEMB sub-TLV

The FAEMB sub-TLV MUST appear only once in the FAD sub-TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver. The Maximum Link Bandwidth as advertised in Extended Link TLV in the Extended Link Opaque LSA in OSPFv2 [RFC7684] or as a sub-TLV of the Router-Link TLV in the E-Router-LSA Router-Link TLV in OSPFv3 [RFC8362] MUST be compared against the Minimum bandwidth advertised in FAEMB sub-TLV. If the link bandwidth is lower than the Minimum bandwidth advertised in FAEMB sub-TLV, the link MUST be excluded from the Flex-Algorithm topology. If a link does not have the Maximum Link Bandwidth advertised but the FAD contains this sub-TLV, then that link MUST be included in the topology and proceed to apply further pruning rules for the link.

## 3.2.2. OSPF Exclude Maximum Delay sub-TLV

OSPF Flex-Algorithm Exclude Maximum Delay sub-TLV (FAEMD) is a sub-TLV of the OSPF FAD TLV. It has the following format.



where:

Type: TBD

Length: 3 octets

Max link delay: Maximum link delay in microseconds

Figure 6: OSPF FAEMD sub-TLV

The FAEMD sub-TLV MUST appear only once in the OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver. The Min Unidirectional Link Delay as advertised by sub-sub-TLV 12 of ASLA sub-TLV [RFC 8920], MUST be compared against the Maximum delay advertised in FAEMD sub-TLV. If the Min Unidirectional Link Delay is higher than the Maximum delay advertised in FAEMD sub-TLV, the link MUST be excluded from the Flex-Algorithm topology. If a link does not have the Min Unidirectional Link Delay advertised but the FAD contains this sub-TLV, then that link MUST NOT be excluded from the topology based on the Maximum Delay constraint.

## 4. Bandwidth Metric Advertisement

Historically, IGP implementations have made default metric assignments based on link bandwidth. This has proven to be useful, but has suffered from having different defaults across implementations and from the rapid growth of link bandwidths. With Flex-Algorithm, the network administrator can define a function that will produce a metric for each link have each node automatically compute each link's metric based its bandwidth.

This document defines a new standard metric type for this purpose called the "Bandwidth Metric". The Bandwidth Metric MAY be advertised in the Generic Metric sub-TLV with the metric type set to "Bandwidth Metric". ISIS and OSPF will advertise this new type of metric in their link advertisements. Bandwidth metric is a link

attribute and for advertisement and processing of this attribute for Flex-algorithm purposes, MUST follow the the section 12 of [I-D.ietf-lsr-flex-algo]

Flex-Algorithm uses this metric type by specifying the bandwidth metric as the metric type in a FAD TLV. A FAD TLV may also specify an automatic computation of the bandwidth metric based on a links advertised bandwidth. An explicit advertisement of a link's bandwidth metric using the Generic Metric sub-TLV overrides this automatic computation. The automatic bandwidth metric calculation sub-TLVs are advertised in FAD TLV and these parameters are applicable to applications such as Flex-algorithm that make use of the FAD TLV.

#### 4.1. Automatic Metric Calculation

Networks which are designed to be highly regular and follow uniform metric assignment may want to simplify their operations by automatically calculating the bandwidth metric. When a FAD advertises the metric type as Bandwidth Metric and the link does not have the Bandwidth Metric advertised, automatic metric derivation can be used with additional FAD constraint advertisements as described in this section.

If a link's bandwidth changes, then the delay in learning about the change may create the possibility of micro-loops in the topology. This is no different from the IGP's susceptibility to micro-loops during a metric change. The micro-loop avoidance procedures described in [I-D.bashandy-rtgwg-segment-routing-uloop] can be used to avoid micro-loops when the automatic metric calculation is deployed.

Computing the metric between adjacent systems based on bandwidth becomes more complex in the face of parallel adjacencies. If there are parallel adjacencies between systems, then the bandwidth between the systems is the sum of the bandwidth of the parallel links. This is somewhat more complex to deal with, so there is an optional mode for computing the aggregate bandwidth.

##### 4.1.1. Automatic Metric Calculation Modes

#### 4.1.1.1. Simple Mode

In simple mode, the Maximum Link Bandwidth of a single Layer 3 link is used to derive the metric. This mode is suitable for deployments that do not use parallel Layer 3 links. In this case, the computation of the metric is straightforward. If a layer 3 link is composed of a layer 2 bundle, then the link bandwidth is the sum of the bandwidths of the working components and may vary with layer 2 link failures.

#### 4.1.1.2. Interface Group Mode

The simple mode of metric calculation may not work well when there are multiple parallel layer 3 interfaces between two nodes. Ideally, the metric between two systems should be the same given the same bandwidth, whether the bandwidth is provided by parallel layer 2 links or parallel layer 3 links. To address this, in Interface Group Mode, nodes MUST compute the aggregate bandwidth of all parallel adjacencies, MUST derive the metric based on the aggregate bandwidth, and MUST apply the resulting metric to each of the parallel adjacencies.

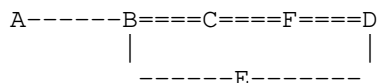


Figure 7: Parallel interfaces

For example, in the above diagram, there are two parallel links between B->C, C->F, F->D. Let us assume the link bandwidth is uniform 10Gbps on all links and the metric for each link will be the same. Traffic from B to D will be forwarded B->E->D. Since the bandwidth is higher on the B->C->F->D path, the metric for that path should be lower, and that path should be selected. Interface Group Mode is preferred in cases where there are parallel layer 3 links.

In the interface group mode, every node MUST identify the set of parallel links between a pair of nodes based on IGP link advertisements and MUST consider cumulative bandwidth of the parallel links while arriving at the metric of each link.

#### 4.1.2. Automatic Metric Calculation Methods

In automatic metric calculation for simple and interface group mode, Maximum Link Bandwidth of the links is used to derive the metric. There are two types of automatic metric derivation methods.

##### 1. Reference bandwidth method

## 2. Bandwidth thresholds method

### 4.1.2.1. Reference Bandwidth method

In many networks, the metric is inversely proportional to the link bandwidth. The administrator or implementation selects a reference bandwidth and the metric is derived by dividing the reference bandwidth by the advertised Maximum Link Bandwidth. Advertising the reference bandwidth in the FAD constraints allows the metric computation to be done automatically. Centralized control of this reference bandwidth simplifies management in the case that the reference bandwidth changes. In order to ensure that small bandwidth changes do not change the link metric, it is useful to define the granularity of the bandwidth that is of interest. The link bandwidth will be truncated to this granularity before deriving the metric.

For example,

reference bandwidth = 1000G

Granularity = 20G

The derived metric is 10 for link bandwidth in the range 100G to 119G

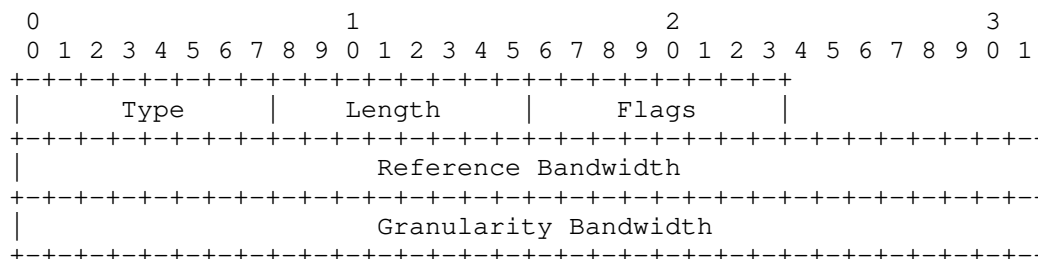
### 4.1.2.2. Bandwidth Thresholds method

The reference bandwidth approach described above provides a uniform metric value for a range of link bandwidths. In certain cases there may be a need to define non-proportional metric values for the varying ranges of link bandwidth. For example, bandwidths from 10G to 30G are assigned metric value 100, bandwidth from 30G to 70G get a metric value of 50, and bandwidths greater than 70G have a metric of 10. In order to support this, a staircase mapping based on bandwidth thresholds is supported in the FAD. This advertisement contains a set of threshold values and associated metrics.

### 4.1.3. ISIS FAD constraint sub-TLVs for automatic metric calculation

#### 4.1.3.1. Reference Bandwidth sub-TLV

This section provides FAD constraint advertisement details for the reference bandwidth method of metric calculation as described in Section 4.1.2.1. The Flexible Algorithm Definition Reference Bandwidth Sub-TLV (FADRB Sub-TLV) is a Sub-TLV of the ISIS FAD sub-TLV. It has the following format:



where:

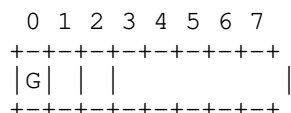
Type: TBD

Length: 9 octets.

Reference Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.

Granularity Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.

Flags:



G-flag: when set, interface group Mode MUST be used to derive total link bandwidth.

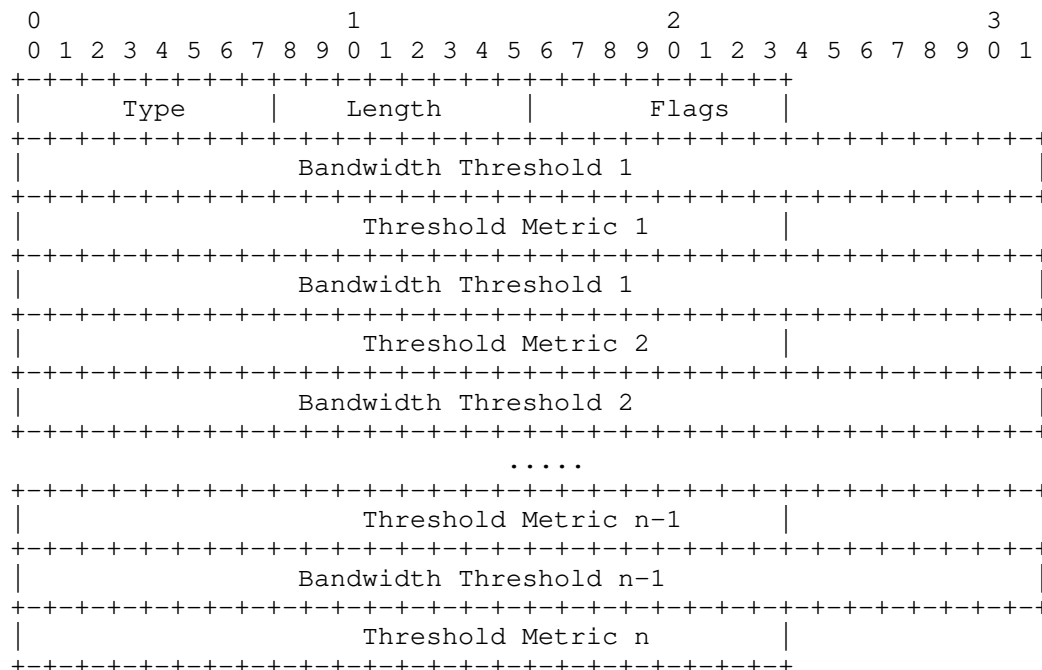
Metric calculation: 
$$\frac{(\text{Reference\_bandwidth})}{(\text{Total\_link\_bandwidth} - (\text{Modulus of}(\text{Total\_link\_bandwidth}, \text{granularity\_bw})) )}$$

Figure 8: ISIS FADRB sub-TLV

Granularity Bandwidth value ensures that the metric does not change when there is a small change in the link bandwidth. The ISIS FADRB Sub-TLV MUST NOT appear more than once in an ISIS FAD sub-TLV. If it appears more than once, the ISIS FAD sub-TLV MUST be ignored by the receiver. If a Generic Metric sub-TLV with Bandwidth metric type is advertised for a link, the Flex-Algorithm calculation MUST use the advertised Bandwidth Metric, and MUST NOT use the automatically derived metric for that link.

## 4.1.3.2. Bandwidth Thresholds sub-TLV

This section provides FAD constraint advertisement details for the Bandwidth Thresholds method of metric calculation as described in Section 4.1.2.2. The Flexible Algorithm Definition Bandwidth Threshold Sub-TLV (FADBT Sub-TLV) is a Sub-TLV of the ISIS FAD sub-TLV. It has the following format:



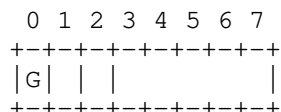
where:

Type: TBD

Length: 1 + n\*7 octets. Here n is equal to number of Threshold Metrics specified.

n MUST be greater than or equal to 1.

Flags:



G-flag: when set, interface group Mode MUST be used to derive total link bandwidth.

```

Staircase bandwidth threshold and associated metric values.
Bandwidth Threshold 1: Minimum Link Bandwidth is encoded in 32 bits in I
EEE
floating point format. The units are bytes per second
.
Bandwidth Threshold 2: Maximum Link Bandwidth is encoded in 32 bits in I
EEE
floating point format. The units are bytes per second
.
Threshold Metric 1 : metric value range (1 - 4,261,412,864)

```

Figure 9: ISIS FADBT sub-TLV

When G-flag is set, the cumulative bandwidth of the parallel links is computed as described in section Section 4.1.1.2. If G-flag is not set, the advertised Maximum Link Bandwidth is used.

When the computed link bandwidth is less than Bandwidth Threshold 1, the MAX\_METRIC value of 4,261,412,864 MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

When the computed link bandwidth is greater than or equal to Bandwidth Threshold 1 and less than Bandwidth Threshold 1, Threshold Metric 1 MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

Similarly, when the computed link bandwidth is greater than or equal to Bandwidth Threshold 1 and less than Bandwidth Threshold 2, Threshold Metric 2 MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

In general, when the computed link bandwidth is greater than or equal to Bandwidth Threshold X AND less than Bandwidth Threshold X+1, Threshold Metric X MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

Finally, when the computed link bandwidth is greater than or equal to Bandwidth Threshold n, then Threshold Metric n MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

The ISIS FADBT Sub-TLV MUST NOT appear more than once in an ISIS FAD sub-TLV. If it appears more than once, the ISIS FAD sub-TLV MUST stop participating in such flex-algorithm.

A FAD MUST NOT contain both FADBT sub-TLV and FADRB sub-TLV. If both these sub-TLVs are advertised in the same FAD for a Flexible Algorithm, the FAD MUST be ignored by the receiver.

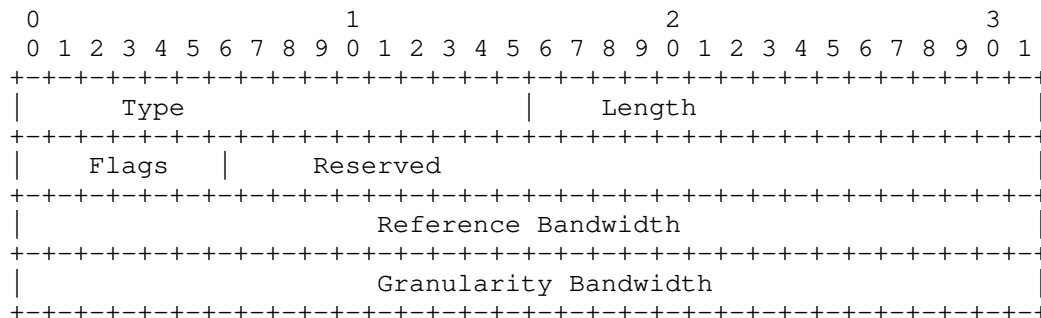


If a Generic Metric sub-TLV with Bandwidth metric type is advertised for a link, the Flex-Algorithm calculation MUST use the Bandwidth Metric advertised on the link, and MUST NOT use the automatically derived metric for that link.

#### 4.1.4. OSPF FAD constraint sub-TLVs for automatic metric calculation

##### 4.1.4.1. Reference Bandwidth sub-TLV

The Flexible Algorithm Definition Reference Bandwidth Sub-TLV (FADRB Sub-TLV) is a Sub-TLV of the OSPF FAD TLV. It has the following format:



where:

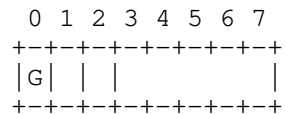
Type: TBD

Length: 14 octets.

Reference Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.

Granularity Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.

Flags:



G-flag: when set, interface group Mode MUST be used to derive total link bandwidth.

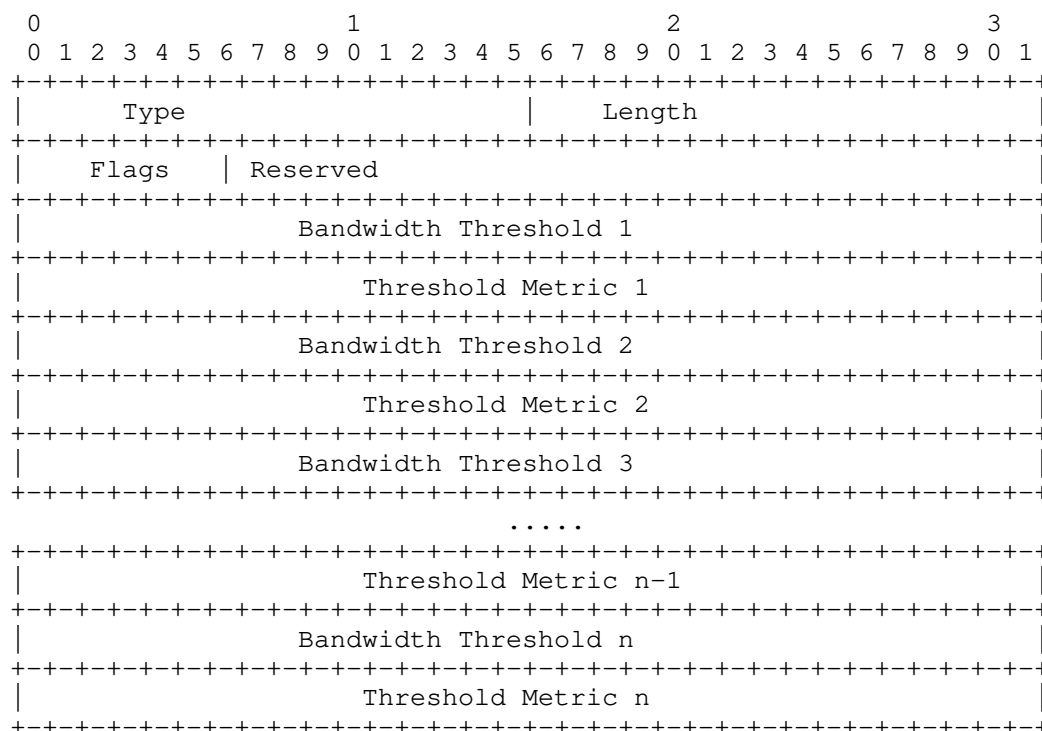
Metric calculation: 
$$\frac{(\text{Reference\_bandwidth})}{(\text{Total\_link\_bandwidth} - (\text{Modulus of}(\text{Total\_link\_bandwidth}, \text{Granularity\_bw}))})}$$

Figure 10: OSPF FADRB sub-TLV

Granularity Bandwidth value is used to ensure that the metric does not change when there is a small change in the link bandwidth. The OSPF FADRB Sub-TLV MUST NOT appear more than once in an OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver. If a Generic Metric sub-TLV with Bandwidth metric type is advertised for a link, the Flex-Algorithm calculation MUST use the advertised Bandwidth Metric on the link, and MUST NOT use the automatically derived metric for that link.

#### 4.1.4.2. Bandwidth Threshold sub-TLV

The Flexible Algorithm Definition Bandwidth Thresholds Sub-TLV (FADBT Sub-TLV) is a Sub-TLV of the OSPF FAD TLV. It has the following format:



where:

Type: TBD

Length:  $2 + n \times 8$  octets. Here  $n$  is equal to number of Threshold Metrics specified.

n MUST be greater than or equal to 1.

Flags:

```

      0 1 2 3 4 5 6 7
    +--+--+--+--+--+--+
    |G|  |  |  |  |  |
    +--+--+--+--+--+--+

```

G-flag: when set, interface group Mode MUST be used to derive total link bandwidth.

Staircase bandwidth threshold and associated metric values.

Bandwidth Threshold 1: Minimum Link Bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second

Bandwidth Threshold 2: Maximum Link Bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second

Threshold Metric 1 : metric value range (1 - 4,294,967,296)

Figure 11: OSPF FADBT sub-TLV

When G-flag is set, the cumulative bandwidth of the parallel links is computed as described in section Section 4.1.1.2. If G-flag is not set, the advertised Maximum Link Bandwidth is used.

When the computed link bandwidth is less than Bandwidth Threshold 1 , the MAX\_METRIC value of 4,294,967,296 MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

When the computed link bandwidth is greater than or equal to Bandwidth Threshold 1 and less than Bandwidth Threshold 2, Threshold Metric 1 MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

Similarly, when the computed link bandwidth is greater than or equal to Bandwidth Threshold 2 and less than Bandwidth Threshold 3, Threshold Metric 2 MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

In general, when the computed link bandwidth is greater than or equal to Bandwidth Threshold X AND less than Bandwidth Threshold X+1, Threshold Metric X MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

Finally, when the computed link bandwidth is greater than or equal to Bandwidth Threshold  $n$ , then Threshold Metric  $n$  MUST be assigned as the Bandwidth Metric on the link during Flex-Algorithm SPF calculation.

The ISIS FADBT Sub-TLV MUST NOT appear more than once in an ISIS FAD sub-TLV. If it appears more than once, the ISIS FAD sub-TLV MUST stop participating in such flex-algorithm.

A FAD MUST NOT contain both FADBT sub-TLV and FADRB sub-TLV. If both these sub-TLVs are advertised in the same FAD for a Flexible Algorithm, the FAD MUST be ignored by the receiver.

If a Generic Metric sub-TLV with Bandwidth metric type is advertised for a link, the Flex-Algorithm calculation MUST use the Bandwidth Metric advertised on the link, and MUST NOT use the automatically derived metric for that link.

## 5. Bandwidth metric considerations

This section specifies the rules of deriving the Bandwidth Metric if and only if the winning FAD for the Flex-Algorithm specifies the metric-type as "Bandwidth Metric".

1. If the Generic Metric sub-TLV with Bandwidth metric type is advertised for the link as described in Section 4, it MUST be used during the Flex-Algorithm calculation.
2. If the Generic Metric sub-TLV with Bandwidth metric type is not advertised for the link and the winning FAD for the Flex-Algorithm does not specify the automatic bandwidth metric calculation (as defined in Section 4.1), the Bandwidth Metric is considered as not being advertised for the link.
3. If the Generic Metric sub-TLV with Bandwidth metric type is not advertised for the link and the winning FAD for the Flex-Algorithm specifies the automatic bandwidth metric calculation (as defined in Section 4.1), the Bandwidth Metric metric MUST be automatically calculated as per the procedures defined in Section 4.1. If the Bandwidth Metric can not be calculated due to lack of Flex-Algorithm specific ASLA advertisement of sub-sub-TLV 9 [RFC 8919], or in case of IS-IS, in presence of the L-Flag in the Flex-Algorithm specific ASLA advertisement the lack of sub-TLV 9 in the TLV 22/222/23/223/141 [RFC 5305], the Bandwidth Metric is considered as not being advertised for the link.

## 6. Calculation of Flex-Algorithm paths

Two new additional rules are added to the existing rules in the Flex-rules specified in sec 13 of [I-D.ietf-lsr-flex-algo].

6. Check if any exclude FAEMB rule is part of the Flex-Algorithm definition. If such exclude rule exists and the link has Maximum Link Bandwidth advertised, check if the link bandwidth satisfies the FAEMB rule. If the link does not satisfy the FAEMB rule, the link MUST be pruned from the computation.

7. Check if any exclude FAEMD rule is part of the Flex-Algorithm definition. If such exclude rule exists and the link has Min Unidirectional link delay advertised, check if the link delay satisfies the FAEMD rule. If the link does not satisfy the FAEMD rule, the link MUST be pruned from the computation.

## 7. Backward Compatibility

## 8. Security Considerations

TBD

## 9. IANA Considerations

### 9.1. IGP Metric-Type Registry

Type: Suggested 3 (TBA)

Description: Bandwidth metric

Reference: This document

Type: 128 to 255(TBA)

Description: User defined metric

Reference: This document

### 9.2. ISIS Sub-Sub-TLVs for Flexible Algorithm Definition Sub-TLV

Type: Suggested 6 (TBA)

Description: ISIS Exclude Minimum Bandwidth sub-TLV

Reference: This document Section 3.1.1

Type: Suggested 7 (TBA)

Description: ISIS Exclude Maximum Delay sub-TLV

Reference: This document Section 3.1.2

Type: Suggested 8 (TBA)

Description: ISIS Reference Bandwidth sub-TLV

Reference: This document Section 4.1.3.1

Type: Suggested 9 (TBA)

Description: ISIS Threshold Metric sub-TLV

Reference: This document Section 4.1.3.2

### 9.3. OSPF Sub-TLVs for Flexible Algorithm Definition Sub-TLV

Type: Suggested 6 (TBA)

Description: OSPF Exclude Minimum Bandwidth sub-TLV

Reference: This document Section 3.2.1

Type: Suggested 7 (TBA)

Description: OSPF Exclude Maximum Delay sub-TLV

Reference: This document Section 3.2.2

Type: Suggested 8 (TBA)

Description: OSPF Reference Bandwidth sub-TLV

Reference: This document Section 4.1.4.1

Type: Suggested 9 (TBA)

Description: OSPF Threshold Metric sub-TLV

Reference: This document Section 4.1.4.2

### 9.4. Sub-TLVs for TLVs 22, 23, 25, 141, 222, and 223

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.1

#### 9.5. OSPFv2 Extended Link TLV Sub-TLVs

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.2

#### 9.6. Types for sub-TLVs of TE Link TLV (Value 2)

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.2

#### 9.7. OSPFv3 Extended-LSA Sub-TLVs

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.2

### 10. Acknowledgements

Many thanks to Chris Bowers, Krzysztof Szarcowitz, Julian Lucek, Ram Santhanakrishnan, Ketan Talaulikar for discussions and inputs.

### 11. Contributors

1. Salih K A

Juniper Networks

salih@juniper.net

### 12. References

#### 12.1. Normative References

[I-D.ietf-lsr-flex-algo]

Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and  
A. Gulko, "IGP Flexible Algorithm", Work in Progress,



Internet-Draft, draft-ietf-lsr-flex-algo-18, 25 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-lsr-flex-algo-18.txt>>.

- [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>>.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <<https://www.rfc-editor.org/info/rfc3630>>.
- [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>>.
- [RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <<https://www.rfc-editor.org/info/rfc7684>>.

## 12.2. Informative References

- [I-D.bashandy-rtgwg-segment-routing-uloop] Bashandy, A., Filsfils, C., Litkowski, S., Decraene, B., Francois, P., and P. Psenak, "Loop avoidance using Segment Routing", Work in Progress, Internet-Draft, draft-bashandy-rtgwg-segment-routing-uloop-12, 22 December 2021, <<https://www.ietf.org/archive/id/draft-bashandy-rtgwg-segment-routing-uloop-12.txt>>.
- [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>>.
- [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>>.

- [RFC7471] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, DOI 10.17487/RFC7471, March 2015, <<https://www.rfc-editor.org/info/rfc7471>>.
- [RFC8570] Ginsberg, L., Ed., Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 8570, DOI 10.17487/RFC8570, March 2019, <<https://www.rfc-editor.org/info/rfc8570>>.

Authors' Addresses

Shraddha Hegde  
Juniper Networks Inc.  
Exora Business Park  
Bangalore 560103  
KA  
India  
Email: [shraddha@juniper.net](mailto:shraddha@juniper.net)

William Britto A J  
Juniper Networks Inc.  
Email: [wbilliam@juniper.net](mailto:wbilliam@juniper.net)

Rajesh Shetty  
Juniper Networks Inc.  
Email: [mrjesh@juniper.net](mailto:mrjesh@juniper.net)

Bruno Decraene  
Orange  
Email: [bruno.decraene@orange.com](mailto:bruno.decraene@orange.com)

Peter Psenak  
Cisco Systems  
Email: [ppsenak@cisco.com](mailto:ppsenak@cisco.com)

Tony Li  
Arista Networks  
Email: [tony.li@tony.li](mailto:tony.li@tony.li)

Network Working Group  
Internet-Draft  
Intended status: Experimental  
Expires: 12 June 2022

A. Przygienda, Ed.  
C. Bowers  
Juniper  
Y. Lee  
A. Sharma  
Comcast  
R. White  
Juniper  
9 December 2021

IS-IS Flood Reflection  
draft-ietf-lsr-isis-flood-reflection-07

Abstract

This document describes a backwards compatible, optional IS-IS extension that allows the creation of IS-IS flood reflection topologies. Flood reflection allows topologies in which L1 areas provide transit forwarding for L2 using all available L1 nodes internally. It accomplishes this by creating L2 flood reflection adjacencies within each L1 area. Those adjacencies are used to flood L2 LSPDUs, and they are used in the L2 SPF computation. However, they are not used for forwarding within the flood reflection cluster. This arrangement gives the L2 topology significantly better scaling properties. As additional benefit, only those routers directly participating in flood reflection have to support the feature. This allows for the incremental deployment of scalable L1 transit areas in an existing network, without the necessity of upgrading other routers in the network.

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 [RFC2119].

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 12 June 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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Glossary . . . . .	8
3. Further Details . . . . .	9
4. Encodings . . . . .	9
4.1. Flood Reflection TLV . . . . .	10
4.2. Flood Reflection Discovery Sub-TLV . . . . .	11
4.3. Flood Reflection Discovery Tunnel Type Sub-Sub-TLV . . . . .	12
4.4. Flood Reflection Adjacency Sub-TLV . . . . .	13
4.5. Flood Reflection Discovery . . . . .	14
4.6. Flood Reflection Adjacency Formation . . . . .	15
5. Route Computation . . . . .	16
5.1. Tunnel Based Deployment . . . . .	16
5.2. No Tunnel Deployment . . . . .	16
6. Redistribution of Prefixes . . . . .	17
7. Special Considerations . . . . .	17
8. IANA Considerations . . . . .	18
8.1. New IS-IS TLV Codepoint . . . . .	18
8.2. Sub TLVs for TLV 242 . . . . .	18
8.3. Sub-sub TLVs for Flood Reflection Discovery sub-TLV . . . . .	18
8.4. Sub TLVs for TLV 22, 23, 25, 141, 222, and 223 . . . . .	18
9. Security Considerations . . . . .	19
10. Acknowledgements . . . . .	19
11. References . . . . .	19
11.1. Informative References . . . . .	19
11.2. Normative References . . . . .	19

Authors' Addresses . . . . . 20

## 1. Introduction

This section introduces the problem space and outlines the solution. Some of the terms may be unfamiliar to reader without extensive IS-IS background and in such case a glossary is provided in Section 2 and can be referenced.

Due to the inherent properties of link-state protocols the number of IS-IS routers within a flooding domain is limited by processing and flooding overhead on each node. While that number can be maximized by well written implementations and techniques such as exponential back-offs, IS-IS will still reach a saturation point where no further routers can be added to a single flooding domain. In some L2 backbone deployment scenarios, this limit presents a significant challenge.

The traditional approach to increasing the scale of an IS-IS deployment is to break it up into multiple L1 flooding domains and a single L2 backbone. This works well for designs where an L2 backbone connects L1 access topologies, but it is limiting where a large L2 is supposed to span large number of routers. In such scenarios, an alternative approach is to consider multiple L2 flooding domains connected together via L1 flooding domains. In other words, L2 flooding domains are connected by "L1/L2 lanes" through the L1 areas to form a single L2 backbone again. Unfortunately, in its simplest implementation, this requires the inclusion of most, or all, of the transit L1 routers as L1/L2 to allow traffic to flow along optimal paths through such transit areas. Consequently, this approach fails to reduce the number of L2 routers involved, so it fails to increase the scalability of the L2 backbone.

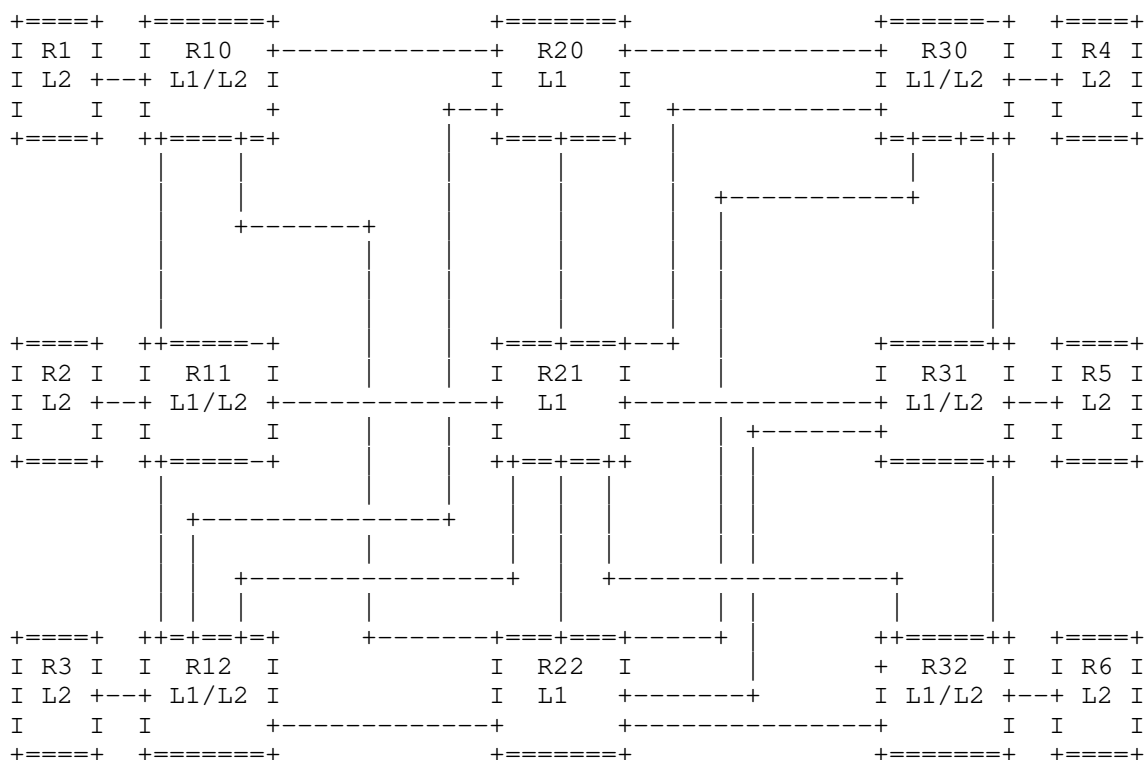


Figure 1: Example Topology of L1 with L2 Borders

Figure 1 is an example of a network where a topologically rich L1 area is used to provide transit between six different L2-only routers (R1-R6). Note that the six L2-only routers do not have connectivity to one another over L2 links. To take advantage of the abundance of paths in the L1 transit area, all the intermediate systems could be placed into both L1 and L2, but this essentially combines the separate L2 flooding domains into a single one, triggering again maximum L2 scale limitation we try to address in first place.

A more effective solution would allow to reduce the number of links and routers exposed in L2, while still utilizing the full L1 topology when forwarding through the network.

[RFC8099] describes Topology Transparent Zones (TTZ) for OSPF. The TTZ mechanism represents a group of OSPF routers as a full mesh of adjacencies between the routers at the edge of the group. A similar mechanism could be applied to IS-IS as well. However, a full mesh of adjacencies between edge routers (or L1/L2 nodes) significantly

limits the scale of the topology. The topology in Figure 1 has 6 L1/L2 nodes. Figure 2 illustrates a full mesh of L2 adjacencies between the 6 L1/L2 nodes, resulting in  $(5 * 6)/2 = 15$  L2 adjacencies. In a somewhat larger topology containing 20 L1/L2 nodes, the number of L2 adjacencies in a full mesh rises to 190.

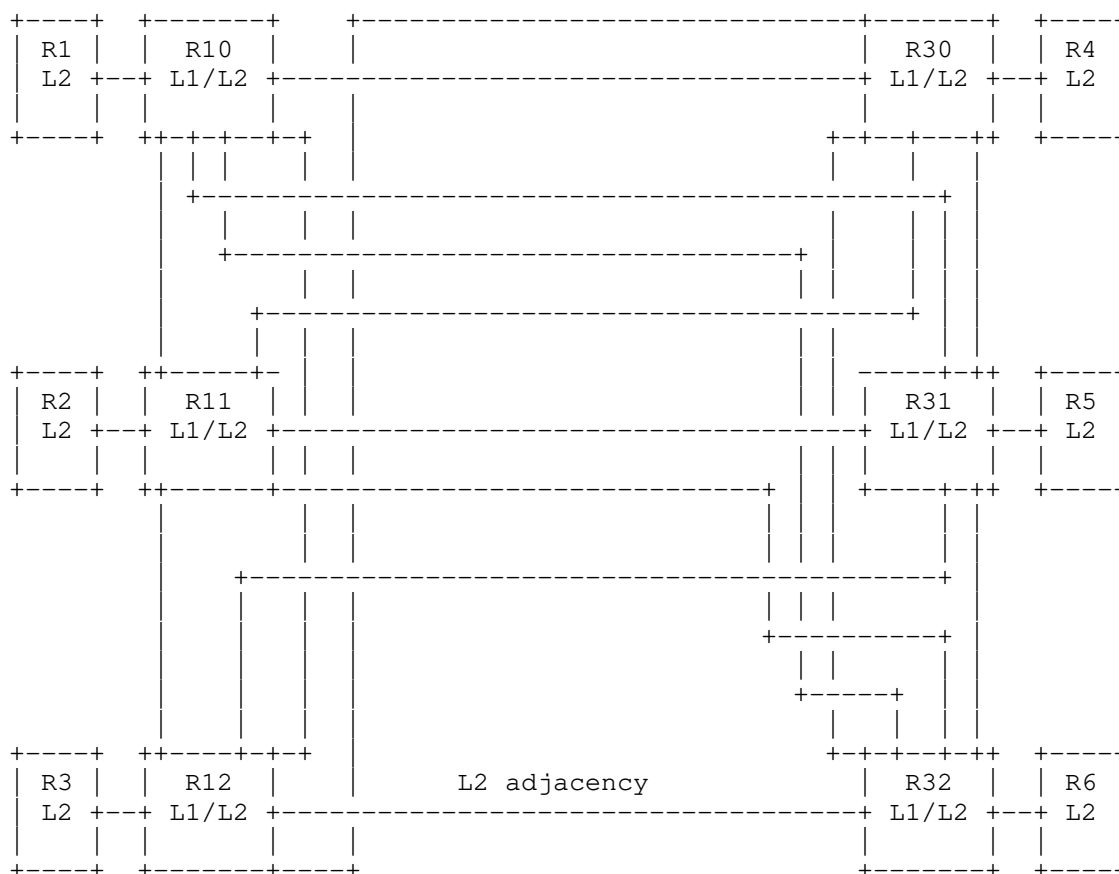


Figure 2: Example topology represented in L2 with a full mesh of L2 adjacencies between L1/L2 nodes

BGP, as specified in [RFC4271], faced a similar scaling problem, which has been solved in many networks by deploying BGP route reflectors [RFC4456]. We note that BGP route reflectors do not necessarily have to be in the forwarding path of the traffic. This incongruity of forwarding and control path for BGP route reflectors allows the control plane to scale independently of the forwarding plane.

We propose here a similar solution for IS-IS. A simple example of what a flood reflector control plane approach would look like is shown in Figure 3, where router R21 plays the role of a flood reflector. Each L1/L2 ingress/egress router builds a tunnel to the flood reflector, and an L2 adjacency is built over each tunnel. In this solution, we need only 6 L2 adjacencies, instead of the 15 needed for a full mesh. In a somewhat larger topology containing 20 L1/L2 nodes, this solution requires only 20 L2 adjacencies, instead of the 190 need for a full mesh. Multiple flood reflectors can be used, allowing the network operator to balance between resilience, path utilization, and state in the control plane. The resulting L2 adjacency scale is  $R*n$ , where  $R$  is the number of flood reflectors used and  $n$  is the number of L1/L2 nodes. This compares quite favorably with  $n*(n-1)/2$  L2 adjacencies required in a topologically fully meshed L2 solution.

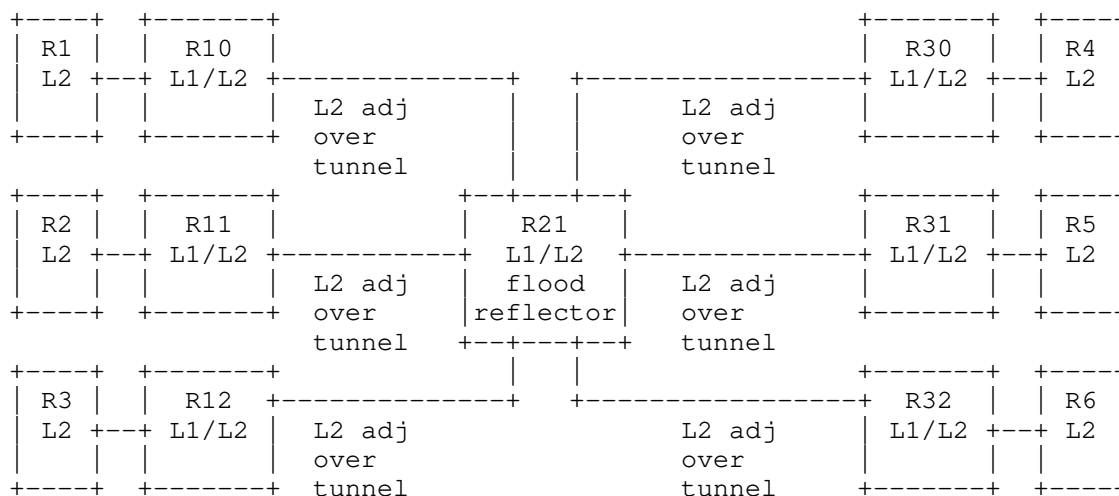


Figure 3: Example topology represented in L2 with L2 adjacencies from each L1/ L2 node to a single flood reflector

As illustrated in Figure 3, when R21 plays the role of flood reflector, it provides L2 connectivity among all of the previously disconnected L2 islands by refloding all L2 LSPDUs. At the same time, R20 and R22 in Figure 1 remain L1-only routers. L1-only routers and L1-only links are not visible in L2. In this manner, the flood reflector allows us provide L2 control plane connectivity in a scalable manner.



As described so far, the solution illustrated in Figure 3 relies only on currently standardized IS-IS functionality. Without new functionality, however, the data traffic will traverse only R21. This will unnecessarily create a bottleneck at R21 since there is still available capacity in the paths crossing the L1-only routers R20 and R22 in Figure 1.

Hence, some new functionality is necessary to allow the L1/L2 edge nodes (R10-12 and R30-32 in Figure 3) to recognize that the L2 adjacency to R21 should not be used for forwarding. The L1/L2 edge nodes should forward traffic that would normally be forwarded over the L2 adjacency to R21 over L1 links instead. This would allow the forwarding within the L1 area to use the L1-only nodes and links shown in Figure 1 as well. It allows networks to be built that use the entire forwarding capacity of the L1 areas, while at the same time introducing control plane scaling benefits provided by L2 flood reflectors.

This document defines all extensions necessary to support flood reflector deployment:

- \* A 'flood reflector adjacency' for all the adjacencies built for the purpose of reflecting flooding information. This allows these 'flood reflectors' to participate in the IS-IS control plane without being used in the forwarding plane. This is a purely local operation on the L1/L2 ingress; it does not require replacing or modifying any routers not involved in the reflection process. Deployment-wise, it is far less tricky to just upgrade the routers involved in flood reflection rather than have a flag day on the whole IS-IS domain.
- \* An (optional) full mesh of tunnels between the L1/L2 routers, ideally load-balancing across all available L1 links. This harnesses all forwarding paths between the L1/L2 edge nodes without injecting unneeded state into the L2 flooding domain or creating 'choke points' at the 'flood reflectors' themselves. The draft is agnostic as to the tunneling technology used but provides enough information for automatic establishment of such tunnels. The discussion of IS-IS adjacency formation and/or liveness discovery on such tunnels is outside the scope of this draft and is largely choice of the underlying implementation. A solution without tunnels is also possible by applying judicious scoping of reachability information between the levels as described in more details later.

- \* Some way to support reflector redundancy, and potentially some way to auto-discover and advertise such adjacencies as flood reflector adjacencies. Such advertisements may allow L2 nodes outside the L1 to perform optimizations in the future based on this information.

## 2. Glossary

This section is introduced with the intention of allowing quick reference in the more detailed parts of the document to terms used

### Flood Reflector:

Node configured to connect L2 only to flood reflector clients and reflect (reflood) IS-IS L2 LSPs amongst them.

### Flood Reflector Client:

Node configured to build flood reflector adjacencies and normal L2 nodes.

### Flood Reflector Adjacency:

IS-IS L2 adjacency limited by one end being client and the other reflector and agreeing on the same Flood Reflector Cluster ID.

### Flood Reflector Cluster:

Collection of clients and flood reflectors configured with the same cluster identifier. Cluster ID value of 0 SHOULD NOT be used since it may be used in the future for special purposes.

### Tunnel Deployment:

Deployment where flood reflector clients build a partial or full mesh of tunnels in L1 to "shortcut" forwarding of L2 traffic through the cluster.

### No Tunnel Deployment:

Deployment where flood reflector clients redistribute L2 reachability into L1 to allow forwarding through the cluster without underlying tunnels.

### Tunnel Endpoint:

An endpoint that allows to establish a bi-directional tunnel carrying ISIS control traffic or alternately serves as origin of such tunnel.

### L1 shortcut:

A tunnel between two clients visible in L1 only that is used as a next-hop, i.e. to carry data traffic in tunnel deployment mode.

### 3. Further Details

Several considerations should be noted in relation to such a flood reflection mechanism.

First, this allows multi-area IS-IS deployments to scale without any major modifications in the IS-IS implementation on most of the nodes deployed in the network. Unmodified (traditional) L2 routers will compute reachability across the transit L1 area using the flood reflector adjacencies.

Second, the flood reflectors are not required to participate in forwarding traffic through the L1 transit area. These flood reflectors can be hosted on virtual devices outside the forwarding topology.

Third, astute readers will realize that flooding reflection may cause the use of suboptimal paths. This is similar to the BGP route reflection suboptimal routing problem described in [ID.draft-ietf-idr-bgp-optimal-route-reflection-28]. The L2 computation determines the egress L1/L2 and with that can create illusions of ECMP where there is none. And in certain scenarios lead to an L1/L2 egress which is not globally optimal. This represents a straightforward instance of the trade-off between the amount of control plane state and the optimal use of paths through the network often encountered when aggregating routing information.

One possible solution to this problem is to expose additional topology information into the L2 flooding domains. In the example network given, links from router 01 to router 02 can be exposed into L2 even when 01 and 02 are participating in flood reflection. This information would allow the L2 nodes to build 'shortcuts' when the L2 flood reflected part of the topology looks more expensive to cross distance wise.

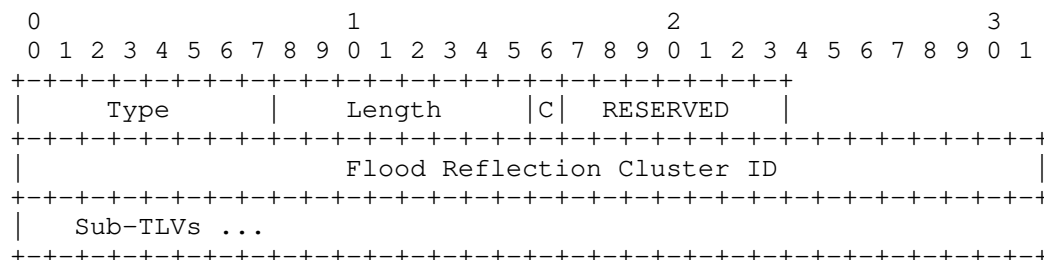
Another possible variation is for an implementation to approximate with the tunnel cost the cost of the underlying topology.

Redundancy can be achieved by building multiple flood reflectors in a L1 area. Multiple flood reflectors do not need any synchronization mechanisms amongst themselves, except standard IS-IS flooding and database maintenance procedures.

### 4. Encodings

#### 4.1. Flood Reflection TLV

The Flood Reflection TLV is a new top-level TLV that MAY appear in L2 IIHs. The Flood Reflection TLV indicates the flood reflector cluster (based on Flood Reflection Cluster ID) that a given router is configured to participate in. It also indicates whether the router is configured to play the role of either flood reflector or flood reflector client. The Flood Reflection Cluster ID and flood reflector roles advertised in the IIHs are used to ensure that flood reflector adjacencies are only formed between a flood reflector and flood reflector client, and that the Flood Reflection Cluster IDs match. The Flood Reflection TLV has the following format:



Type: TBD

Length: The length, in octets, of the following fields.

C (Client): This bit is set to indicate that the router acts as a flood reflector client. When this bit is NOT set, the router acts as a flood reflector. On a given router, the same value of the C-bit MUST be advertised across all interfaces advertising the Flood Reflection TLV in IIHs.

RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Flood Reflection Cluster ID: Flood Reflection Cluster Identifier. These same 32-bit value MUST be assigned to all of the flood reflectors and flood reflector clients in the same L1 area. The value MUST be unique across different L1 areas within the IGP domain. In case of violation of those rules multiple L1 areas may become a single cluster or a single area may split in flood reflection sense and several mechanisms such as auto-discovery of tunnels may not work correctly. On a given router, the same value of the Flood Reflection Cluster ID MUST be advertised across all interfaces advertising the Flood Reflection TLV in IIHs. When a router discovers that a node is using multiple Cluster IDs based

on its advertised TLVs and IIHs, the node MAY adequately log such violations subject to rate limiting. This implies that a flood reflector MUST NOT participate in more than a single L1 area. In case of Cluster ID value of 0, the TLV containing it MUST be ignored.

Sub-TLVs: Optional sub-TLVs. For future extensibility, the format of the Flood Reflection TLV allows for the possibility of including optional sub-TLVs. No sub-TLVs of the Flood Reflection TLV are defined in this document.

The Flood Reflection TLV SHOULD NOT appear more than once in an IIH. A router receiving multiple Flood Reflection TLVs in the same IIH MUST use the values in the first TLV of the lowest numbered fragment and it SHOULD adequately log such violations subject to rate limiting.

#### 4.2. Flood Reflection Discovery Sub-TLV

Flood Reflection Discovery sub-TLV is advertised as a sub-TLV of the IS-IS Router Capability TLV-242, defined in [RFC7981]. The Flood Reflection Discovery sub-TLV is advertised in L1 and L2 LSPs with area flooding scope in order to enable the auto-discovery of flood reflection capabilities. The Flood Reflection Discovery sub-TLV has the following format:

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type										Length										C   Reserved																			
Flood Reflection Cluster ID																																							

Type: TBD

Length: The length, in octets, of the following fields.

C (Client): This bit is set to indicate that the router acts as a flood reflector client. When this bit is NOT set, the router acts as a flood reflector.

RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

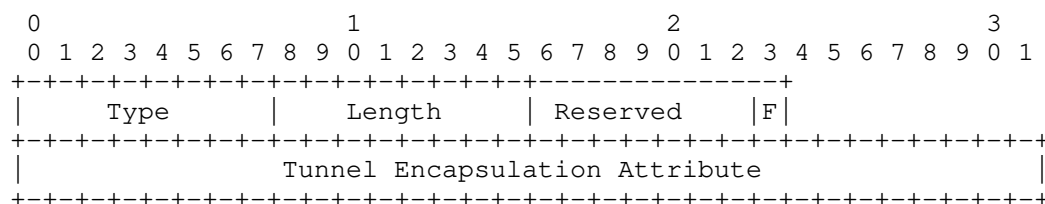
Flood Reflection Cluster ID: The Flood Reflection Cluster Identifier

is the same as that defined in the Flood Reflection TLV and obeys the same rules.

The Flood Reflection Discovery sub-TLV SHOULD NOT appear more than once in TLV 242. A router receiving multiple Flood Reflection Discovery sub-TLVs in TLV 242 MUST use the values in the first sub-TLV of the lowest numbered fragment and it SHOULD adequately log such violations subject to rate limiting.

#### 4.3. Flood Reflection Discovery Tunnel Type Sub-Sub-TLV

Flood Reflection Discovery Tunnel Type sub-sub-TLV is advertised optionally as a sub-sub-TLV of the Flood Reflection Discovery Sub-TLV, defined in Section 4.2. It allows the automatic creation of L2 tunnels to be used as flood reflector adjacencies and L1 shortcut tunnels. The Flood Reflection Tunnel Type sub-sub-TLV has the following format:



Type: TBD

Length: The length, in octets, of zero or more of the following fields.

Reserved: SHOULD be 0 on transmission and ignored on reception.

F Flag: When set indicates flood reflection tunnel endpoint, when clear, indicates possible L1 shortcut tunnel endpoint.

Tunnel Encapsulation Attribute: Carries encapsulation type and further attributes necessary for tunnel establishment as defined in [RFC9012]. Protocol type sub-TLV as defined in [RFC9012] MAY be included but MUST when F flag is set include according type that allows carrying of encapsulated IS-IS frames. Such tunnel type MUST provide according mechanisms to carry up to 'originatingL2LSPBufferSize' sized IS-IS frames across.

A flood reflector receiving Flood Reflection Discovery Tunnel Type sub-sub-TLVs in Flood Reflection Discovery sub-TLV with F flag set SHOULD use one or more of the specified tunnel endpoints to automatically establish one or more tunnels that will serve as flood reflection adjacency(-ies) to the clients advertising the endpoints.

A flood reflection client receiving multiple Flood Reflection Discovery Tunnel Type sub-sub-TLVs in Flood Reflection Discovery sub-TLV with F flag clear from other leaves MAY use one or more of the specified tunnel endpoints to automatically establish one or more tunnels that will serve as L1 tunnel shortcuts to the clients advertising the endpoints.

In case of automatic flood reflection adjacency tunnels and in case IS-IS adjacencies are being formed across L1 shortcuts all the aforementioned rules in Section 4.5 apply as well.

Optional address validation procedures as defined in [RFC9012] MUST be disregarded.

#### 4.4. Flood Reflection Adjacency Sub-TLV

The Flood Reflection Adjacency sub-TLV is advertised as a sub-TLV of TLVs 22, 23, 25, 141, 222, and 223. Its presence indicates that a given adjacency is a flood reflector adjacency. It is included in L2 area scope flooded LSPs. Flood Reflection Adjacency sub-TLV has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |C|  Reserved  |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Flood Reflection Cluster ID
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Type: TBD

Length: The length, in octets, of the following fields.

C (Client): This bit is set to indicate that the router advertising this adjacency is a flood reflector client. When this bit is NOT set, the router advertising this adjacency is a flood reflector.

RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Flood Reflection Cluster ID: The Flood Reflection Cluster Identifier is the same as that defined in the Flood Reflection TLV and obeys the same rules.

The Flood Reflection Adjacency sub-TLV SHOULD NOT appear more than once in a given TLV. A router receiving multiple Flood Reflection Adjacency sub-TLVs in a TLV MUST use the values in the first sub-TLV of the lowest numbered fragment and it SHOULD adequately log such violations subject to rate limiting.

#### 4.5. Flood Reflection Discovery

A router participating in flood reflection as client or reflector MUST be configured as an L1/L2 router. It SHOULD originate the Flood Reflection Discovery sub-TLV with area flooding scope in L1 and L2. Normally, all routers on the edge of the L1 area (those having traditional L2 adjacencies) will advertise themselves as route reflector clients. Therefore, a flood reflector client will have both traditional L2 adjacencies and flood reflector L2 adjacencies.

A router acting as a flood reflector MUST NOT have any traditional L2 adjacencies except with flood reflector clients. It will be an L1/L2 router only by virtue of having flood reflector L2 adjacencies. A router desiring to act as a flood reflector SHOULD advertise itself as such using the Flood Reflection Discovery sub-TLV in L1 and L2.

A given flood reflector or flood reflector client can only participate in a single cluster, as determined by the value of its Flood Reflection Cluster ID and should disregard other routers' TLVs for flood reflection purposes if the cluster ID is not matching.

Upon reception of Flood Reflection Discovery sub-TLVs, a router acting as flood reflector client SHOULD initiate a tunnel towards each flood reflector with which it shares an Flood Reflection Cluster ID using one or more of the tunnel encapsulations provided with F flag being set. The L2 adjacencies formed over such tunnels MUST be marked as flood reflector adjacencies. If the client or reflector has a direct L2 adjacency with the according remote side it SHOULD use it instead of instantiating a new tunnel.

In absence of auto-discovery an implementation MAY use statically configured tunnels to create flood reflection adjacencies.

The IS-IS metrics for all flood reflection adjacencies in a cluster SHOULD be uniform.



Upon reception of Flood Reflection Discover TLVs, a router acting as a flood reflector client MAY initiate tunnels with L1-only adjacencies towards any of the other flood reflector clients with lower router IDs in its cluster using encapsulations with F flag clear. These tunnels MAY be used for forwarding to improve the load-balancing characteristics of the L1 area. If the clients have a direct L2 adjacency they SHOULD use it instead of instantiating a new tunnel.

#### 4.6. Flood Reflection Adjacency Formation

In order to simplify both implementations and network deployments, this draft does not allow the formation of complex hierarchies of flood reflectors and clients or allow multiple clusters in a single L1 area. Consequently, all flood reflectors and flood reflector clients in the same L1 area MUST share the same Flood Reflector Cluster ID. Deployment of multiple cluster IDs in the same L1 area are outside the scope of this document.

A flood reflector MUST only form flood reflection adjacencies with flood reflector clients with matching Cluster ID. A flood reflector MUST NOT form any traditional L2 adjacencies.

Flood reflector clients MUST only form flood reflection adjacencies with flood reflectors with matching Cluster ID.

Flood reflector clients MAY form traditional L2 adjacencies with flood reflector clients or nodes not participating in flood reflection. When two clients form traditional L2 adjacency Cluster ID is disregarded.

The Flood Reflector Cluster ID and flood reflector roles advertised in the Flood Reflection TLVs in IIHs are used to ensure that flood reflection adjacencies that are established meet the above criteria.

On change in either flood reflection role or cluster ID on IIH on the local or remote side the adjacency has to be reset and re-established if possible.

Once a flood reflection adjacency is established, the flood reflector and the flood reflector client MUST advertise the adjacency by including the Flood Reflection Adjacency Sub-TLV in the Extended IS reachability TLV or MT-ISN TLV.

## 5. Route Computation

To ensure loop-free routing, the route reflection client **MUST** follow the normal L2 computation to determine L2 routes. This is because nodes outside the L1 area will generally not be aware that flood reflection is being performed. The flood reflection clients need to produce the same result for the L2 route computation as a router not participating in flood reflection.

### 5.1. Tunnel Based Deployment

In tunnel based option the reflection client, after L2 and L1 computation, **MUST** examine all L2 routes and replace all flood reflector adjacencies with the correct underlying tunnel next-hop to the egress.

### 5.2. No Tunnel Deployment

In case of deployment without underlying tunnels, the necessary L2 routes are distributed into the area, normally as L2->L1 routes. Due to the rules in Section 4.6 the computation in the resulting topology is relatively simple, the L2 SPF from a flood reflector client is guaranteed to reach within a hop the Flood Reflector and in the following hop the L2 egress to which it has a forwarding tunnel again. All the flood reflector tunnel nexthops in the according L2 route can hence be removed and if the L2 route has no other ECMP L2 nexthops, the L2 route **MUST** be suppressed in the RIB by some means to allow the less preferred L2->L1 route to be used to forward traffic towards the advertising egress.

In the particular case the client has L2 routes which are not route reflected, those will be naturally preferred (such routes normally "hot-potato" route of the L1 area). However in the case the L2 route through the flood reflector egress is "shorter" than such present non flood reflected L2 routes, the node **SHOULD** ensure that such routes are suppressed so the L2->L1 towards the egress still takes preference. Observe that operationally this can be resolved in a relatively simple way by configuring flood reflector adjacencies to have a high metric, i.e. the flood reflector topology becomes "last resort" and the leaves will try to "hot-potato" out the area as fast as possible which is normally the desirable behavior.

In deployment scenarios where tunnels are not used, all L1/L2 edge nodes **MUST** be ultimately flood reflector clients except during transition phase.

## 6. Redistribution of Prefixes

When L2 prefixes need to be redistributed into L1 by the route reflector clients a client that does not have any L2 flood reflector adjacencies **MUST NOT** redistribute those routes into the area in case of application of Section 5.2. The L2 prefixes advertisements redistributed into L1 with flood reflectors **SHOULD** be normally limited to L2 intra-area routes (as defined in [RFC7775]), if the information exists to distinguish them from other other L2 prefix advertisements.

On the other hand, in topologies that make use of flood reflection to hide the structure of L1 areas while still providing transit forwarding across them using tunnels, we generally do not need to redistribute L1 prefixes advertisements into L2.

## 7. Special Considerations

In pathological cases setting the overload bit in L1 (but not in L2) can partition L1 forwarding, while allowing L2 reachability through flood reflector adjacencies to exist. In such a case a node cannot replace a route through a flood reflector adjacency with a L1 shortcut and the client can use the L2 tunnel to the flood reflector for forwarding while it **MUST** initiate an alarm and declare misconfiguration.

A flood reflector with directly L2 attached prefixes should advertise those in L1 as well since based on preference of L1 routes the clients will not try to use the L2 flood reflector adjacency to route the packet towards them. A very, very corner case is when the flood reflector is reachable via L2 flood reflector adjacency (due to underlying L1 partition) only in which case the client can use the L2 tunnel to the flood reflector for forwarding towards those prefixes while it **MUST** initiate an alarm and declare misconfiguration.

A flood reflector **SHOULD NOT** set the attached bit on its LSPs.

Instead of modifying the computation procedures one could imagine a flood reflector solution where the Flood Reflector would re-advertise the L2 prefixes with a 'third-party' next-hop but that would have less desirable convergence properties than the solution proposed and force a fork-lift of all L2 routers to make sure they disregard such prefixes unless in the same L1 domain as the Flood Reflector.

Depending on pseudo-node choice in case of a broadcast domain with multiple flood reflectors attached this can lead to a partitioned LAN and hence a router discovering such a condition **MUST** initiate an alarm and declare misconfiguration.

## 8. IANA Considerations

This document requests allocation for the following IS-IS TLVs and Sub-TLVs.

### 8.1. New IS-IS TLV Codepoint

This document requests the following IS-IS TLV:

Value Name	IIH	LSP	SNP	Purge
TBD1    Flood Reflection	y	n	n	n

Suggested value for TBD1 is 161.

### 8.2. Sub TLVs for TLV 242

This document request the following registration in the "sub-TLVs for TLV 242" registry.

Type	Description
TBD2	Flood Reflection Discovery

Suggested value for TBD2 is 161.

### 8.3. Sub-sub TLVs for Flood Reflection Discovery sub-TLV

This document request the following registration in the "sub-sub-TLVs for Flood Reflection Discovery sub-TLV" registry.

Type	Description
TBD3	Flood Reflection Discovery Tunnel Encapsulation Attribute

Suggested value for TBD3 is 161.

### 8.4. Sub TLVs for TLV 22, 23, 25, 141, 222, and 223

This document requests the following registration in the "sub-TLVs for TLV 22, 23, 25, 141, 222, and 223" registry.

Type	Description	22	23	25	141	222	223
TBD4	Flood Reflector Adjacency	y	y	n	y	y	y

Suggested value for TBD4 is 161.

## 9. Security Considerations

This document introduces tunnels carrying IS-IS control traffic via tunnels. In case of statically configured tunnels a deployment SHOULD provide enough security protection to prevent malicious attackers from using the tunnel endpoints. For information used to form dynamically discovered tunnels, it SHOULD be protected by the the deployed IS-IS security mechanism preventing malicious nodes from spoofing rogue information on behalf of other members.

## 10. Acknowledgements

The authors thank Shraddha Hegde, Peter Psenak, Acee Lindem, Robert Raszuk and Les Ginsberg for their thorough review and detailed discussions. Thanks are also extended to Michael Richardson for an excellent routing directorate review.

## 11. References

### 11.1. Informative References

- [ID.draft-ietf-idr-bgp-optimal-route-reflection-28]  
Raszuk et al., R., "BGP Optimal Route Reflection", July 2019, <<https://www.ietf.org/id/draft-ietf-idr-bgp-optimal-route-reflection-28.txt>>.
- [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>>.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <<https://www.rfc-editor.org/info/rfc4456>>.
- [RFC8099] Chen, H., Li, R., Retana, A., Yang, Y., and Z. Liu, "OSPF Topology-Transparent Zone", RFC 8099, DOI 10.17487/RFC8099, February 2017, <<https://www.rfc-editor.org/info/rfc8099>>.

### 11.2. 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, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC7775] Ginsberg, L., Litkowski, S., and S. Previdi, "IS-IS Route Preference for Extended IP and IPv6 Reachability", RFC 7775, DOI 10.17487/RFC7775, February 2016, <<https://www.rfc-editor.org/info/rfc7775>>.
- [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>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

#### Authors' Addresses

Tony Przygienda (editor)  
Juniper  
1137 Innovation Way  
Sunnyvale, CA  
United States of America

Email: [prz@juniper.net](mailto:prz@juniper.net)

Chris Bowers  
Juniper  
1137 Innovation Way  
Sunnyvale, CA  
United States of America

Email: [cbowers@juniper.net](mailto:cbowers@juniper.net)

Yiu Lee  
Comcast  
1800 Bishops Gate Blvd  
Mount Laurel, NJ 08054  
United States of America

Email: [Yiu\\_Lee@comcast.com](mailto:Yiu_Lee@comcast.com)

Alankar Sharma  
Comcast  
1800 Bishops Gate Blvd  
Mount Laurel, NJ 08054  
United States of America

Email: Alankar\_Sharma@comcast.com

Russ White  
Juniper  
1137 Innovation Way  
Sunnyvale, CA  
United States of America

Email: russw@juniper.net

Networking Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 8, 2022

P. Psenak, Ed.  
L. Ginsberg  
Cisco Systems  
K. Talaulikar  
Individual Contributor  
March 7, 2022

IS-IS and OSPF Extension for Event Notification  
draft-ppsenak-lsr-igp-event-notification-01

Abstract

Link-state protocols like IS-IS and OSPF have been designed to distribute state information - state of the local adjacencies, state of the local prefix reachability, etc. Each state can have additional attributes associated with it, but all the attributes are only meaningful while the state exists.

This document extends link-state IGPs to distribute event notifications. An event notification has a very limited lifetime. It is rapidly propagated across the network and leaves no state after its lifetime.

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 September 8, 2022.



## Copyright Notice

Copyright (c) 2022 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 . . . . .	3
2. Requirements for Pulse Notification . . . . .	3
3. IS-IS Pulse PDUs . . . . .	4
3.1. IS-IS Flooding Scoped Pulse LSP . . . . .	4
3.2. IS-IS Flooding Scoped Pulse PSNP . . . . .	5
3.3. Flooding Scope Update Process Operation . . . . .	7
3.3.1. FSP-LSP Generation Procedures . . . . .	8
3.3.2. FSP-LSP Acknowledgement Behavior . . . . .	8
4. Use Case: Supporting BGP-PIC at scale . . . . .	8
4.1. Use of Summarization . . . . .	9
4.2. Use of Pulse in combination w Summarization . . . . .	10
4.3. IS-IS Summary Component Reachability Loss Pulse TLV . . . . .	10
5. Handling of the Control Plane Restart and ISSU . . . . .	13
6. OSPF Pulse Notification . . . . .	13
7. OSPFv3 Pulse Notification . . . . .	14
8. IANA Considerations . . . . .	14
8.1. New IS-IS PDU Types . . . . .	14
8.2. Revised sub-TLV table . . . . .	14
8.3. IS-IS Flooding Scope Pulse LSP Entries TLV . . . . .	14
8.4. IS-IS Summary Component Reachability Loss Pulse TLV . . . . .	14
9. Security Considerations for ISIS . . . . .	14
10. Security Considerations for OSPF . . . . .	15
11. Contributors . . . . .	15
12. References . . . . .	15
12.1. Normative References . . . . .	15
12.2. Informative References . . . . .	16
Appendix A. BGP Pulse Handling . . . . .	16
Authors' Addresses . . . . .	17

## 1. Introduction

Link-state IGP protocols like IS-IS and OSPF are primarily used to distribute routing information between routers belonging to a single Autonomous System (AS) and to calculate the reachability for IPv4 or IPv6 prefixes advertised by the individual nodes inside the AS. Each node advertises the state of its local adjacencies, connected prefixes, capabilities, etc. The collection of these states from all the routers inside the area form a link-state database (LSDB) that describes the topology of the area and holds additional state information about the prefixes, router capabilities, etc.

Link-state protocols have been designed to distribute state information. More precisely, it's only the existent or steady state that is being advertised - local adjacencies in UP state, local prefixes that are reachable, etc. When the state does not exist anymore (e.g., adjacency transition to DOWN state), it is simply removed by the advertising node. Each state can have additional attributes associated with it, but all the attributes are only meaningful while the state exists.

There are certain types of events, that do not represent a steady state and therefore cannot be advertised, that may be useful for the network operation.

This document introduces the capability in link-state protocols to propagate event notifications that have a short and limited lifetime and do not introduce a state into IS-IS, OSPFv2, and OSPFv3. These event notifications are referred to as pulses to reflect their short-lived nature. Pulses may be used to advertise many types of events including those that are positive or negative in nature and for which there is no associated state that is to be maintained by link-state protocols.

## 2. Requirements for Pulse Notification

This section describes the basic requirements of the pulse based notification for link-state protocols.

Pulse Processing - processing of the pulse on the router is OPTIONAL. It's the decision of the receiver of the pulse whether the pulse is processed and any action is taken.

Reliability - the distribution of the pulses MUST be reliable.

Separation from Link-State - receiving pulses MUST NOT result in any update of the link-state topology or in any route calculation.

Pulse advertisements MUST NOT be sent using existing protocol link-state messages.

Limited Lifetime - pulses are short-lived. There is no flushing or purging mechanism for pulses. They MUST be destroyed after their flooding procedure is complete.

Limited Retransmissions - pulses MUST be retransmitted, as required by the flooding procedure, only for a limited period.

Not Part of Database Sync - pulses MUST NOT be exchanged as part of the initial or post Graceful Restart database synchronization between adjacent peers.

Relevance to routing protocol - use of pulses is restricted to information known to the routing protocol as part of its normal operation

### 3. IS-IS Pulse PDUs

Two new IS-IS PDUs are introduced for pulse propagation:

Flooding Scoped Pulse LSP (FSP-LSP)

Flooding Scoped Pulse PSNP (FSP-PSNP)

#### 3.1. IS-IS Flooding Scoped Pulse LSP

The format of an IS-IS Flooding Scoped Pulse LSP (FSP-LSP) is similar to the format of the Flooding Scoped LSP defined in [RFC7356], with the following fields being removed:

"Reserved"

"Remaining Lifetime"

"Reserved|LSPDBOL|IS Type"

FSP-LSP supports all flooding scopes defined in [RFC7356].

An FS-Pulse-LSP has the following format:

	No. of octets
Intradomain Routeing Protocol Discriminator	1
Length Indicator	1
Version/Protocol ID Extension	1
ID Length	1
R R R  PDU Type	1
Version	1
P  Scope	1
PDU Length	2
FSP-LSP ID	ID Length + 2
Sequence Number	4
Checksum	2
: Variable Length Fields :	Variable

PDU Type: 7 - (suggested - to be assigned by IANA)

All fields as defined in [RFC7356] for FS-LSPs

### 3.2. IS-IS Flooding Scoped Pulse PSNP

The format of an IS-IS Flooding Scoped Pulse PSNP (FSP-PSNP) is similar to the format of the Flooding Scoped PSNP defined in [RFC7356]

FSP-PSNP supports all flooding scopes defined in [RFC7356].

An FSP-PSNP has the following format:

	No. of octets
Intradomain Routeing Protocol Discriminator	1
Length Indicator	1
Version/Protocol ID Extension	1
ID Length	1
R R R  PDU Type	1
Version	1
Reserved	1
U  Scope	1
PDU Length	2
Source ID	ID Length + 1
: Variable Length Fields :	Variable

PDU Type: 8 (Suggested - to be assigned by IANA) defined in [ISO10589].

All fields of the FSP-PSNP match the definition from Flooding Scoped PSNP in [RFC7356].

Variable-length fields - list of TLVs.

This document defines a new TLV to be included in FSP-PSNPs: Flooding Scope Pulse LSP Entries TLV (FSP-LSP Entries TLV) that has the following format:

	No. of octets
Type	1
Length	1
FSP-LSP ID	ID Length + 2
Sequence Number	4
Checksum	2
:	:
FSP-LSP ID	ID Length + 2
Sequence Number	4
Checksum	2

Type: 29 (Suggested - to be assigned by IANA)

Length: (ID Length + 8) \* number of entries

FSP-LSP ID: The ID of the FSP-LSP being acknowledged

Sequence Number: Sequence number of the FSP-LSP being acknowledged

Checksum: Checksum reported in the FSP-LSP

### 3.3. Flooding Scope Update Process Operation

The Update Process in [ISO10589] is responsible for reliable flooding of LSPs. In the case of FSP-LSPs, the lack of persistence introduces some changes in how the Update Process operates.

Analagous to what is defined in [RFC7356], there is a separate instance of the Update Process for each scope supported for FSP-LSPs. The circuit(s) on which FSP-LSPs are flooded is limited to those circuits that are participating in the given scope. Consistent support of a given flooding scope on a circuit by all routers operating on that circuit is required.

FSP-LSPs are not meant to be retained beyond the minimum time needed to process the information and to provide a reasonable opportunity for flooding the information to neighbors. FSP-LSPs are also not

synchronized on adjacency establishment and/or Graceful Restart [RFC8706]. For this reason, an FSP Complete Sequence Number PDU is NOT REQUIRED. Flooding of an FSP-LSP on a circuit ceases after a configurable number of retries. Default number of retries is RECOMMENDED to be 3.

Receipt of an FSP-PSNP with a matching Flooding Scope Pulse LSP Entry serves as an acknowledgment of receipt of an FSP-LSP on that circuit.

FSP-LSPs SHOULD be retained in the FSP Scope Specific LSDB for ZeroAgeLifetime (60 seconds). This is done to support reliable flooding of the FSP-LSP and to minimize the possibility of reprocessing a previously received FSP-LSP.

### 3.3.1. FSP-LSP Generation Procedures

Although sequence numbers in FSP-LSPs are less important than in traditional LSPs since FSP-LSPs are not retained for a significant period and are not purged, they are still useful to identity a newer version of a given FSP-LSP. Nodes which originate FSP-LSPs MUST remember the last sequence number used for a given FSP-LSP and increment the sequence number when generating a new version.

FSP-LSP generation SHOULD utilize the "next" FSP-LSP ID each time new pulse information needs to be advertised i.e., if the most recent FSP-LSP ID used was A-00.n, the next set of pulse information SHOULD be advertised using FSP-LSP.ID A-00.n+1. This minimizes the possibility of confusion if other routers in the network have not yet removed A-00.n from their LSPDB.

### 3.3.2. FSP-LSP Acknowledgement Behavior

Determining whether a received FSP-LSP is newer than a previously received copy follows the rules defined for LSPs defined in [ISO10589].

Received FSP-LSPs which are either newer or the same as an existing entry in the LSPDB are acknowledged using FSP-PSNPs.

Received FSP-LSPs which are older than existing entries in the LSPDB are ignored. The sender of the FSP-LSP will in any case cease flooding such an FSP-LSP after a modest number of retries.

## 4. Use Case: Supporting BGP-PIC at scale

In this section we present one practical use case of event based notification in link-state routing protocols.

The growth of networks running a link-state routing protocol results in the addition of more state that presents itself in the form of scalability and convergence challenges. The organization of networks into levels/areas and IGP domains helps limit the scope of link-state information within certain boundaries. However, the state related to prefix reachability often requires propagation across a multi-area/level and/or multi-domain IGP network.

Techniques such as summarization have been used traditionally to address the scale challenges associated with advertising prefix state outside of local area/domain.

However, this results in suppression of the individual prefix state that is useful for triggering fast-convergence mechanisms outside of the IGP's - e.g., BGP PIC Edge [I-D.ietf-rtgwg-bgp-pic].

In such a scenario, it is desirable to enable the notification of events, such as an individual prefix becoming unreachable, outside of the local area/domain and across the network in a manner that does not leave behind any persistent state in the link-state database.

#### 4.1. Use of Summarization

Deployment of large networks may utilize a significant number of discrete IGP areas. Advertisement of inter-area prefixes is limited to summaries to reduce the number of prefix advertisements which need to be flooded domain-wide.

Consider a network consisting of 100 areas with 1K prefixes/area. In the absence of summarization, there are 100 K prefixes which would be advertised domain-wide. If in each area there are two Area Border Routers (ABRs) - for redundancy purposes - each of which is advertising 1K intra-area prefixes into other areas, there would then be  $100 * 2 * 1K = 200K$  prefix advertisements sent domain-wide.

If a single summary address is used to represent reachability for the 1K prefixes within an area, the number of prefix advertisements flooded domain-wide becomes  $100 * 2 * 1 = 200$  summary prefix advertisements.

The use of summarization dramatically reduces the scale of network-wide flooding, but it also means that a change in reachability to any specific destination covered by a summary is not known to routers outside a given area.



#### 4.2. Use of Pulse in combination w Summarization

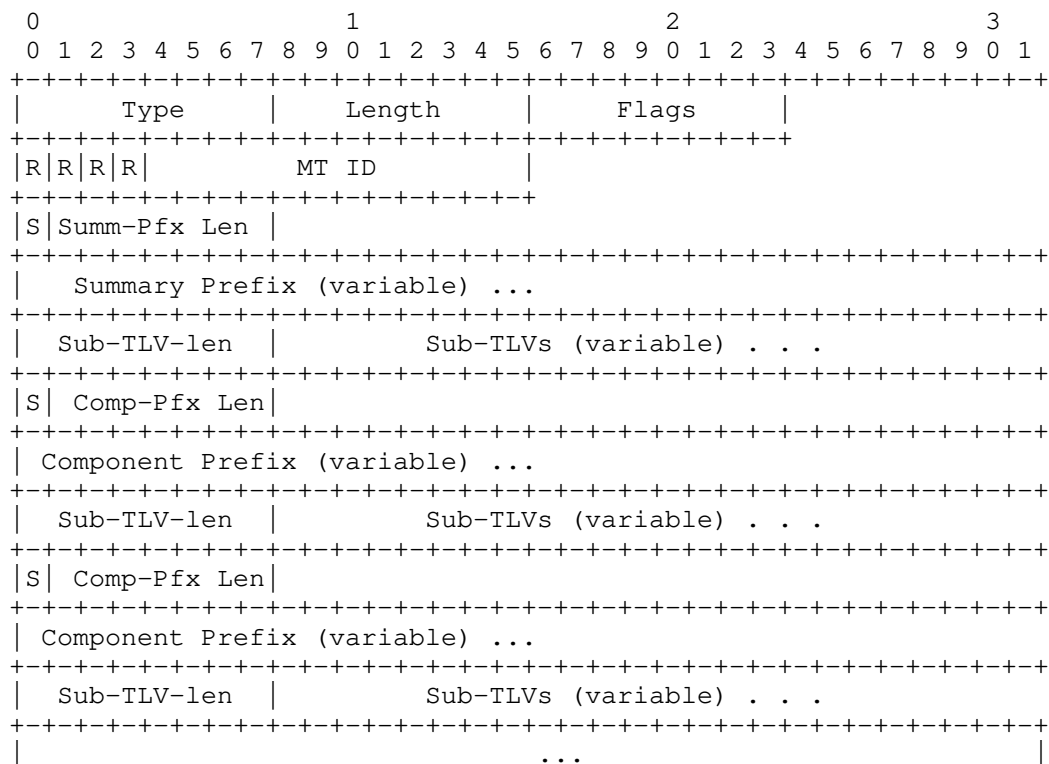
Pulse can be used to signal loss of reachability to an individual destination covered by a summary. In the event that a single node becomes unreachable, this would result in the flooding of 2 pulses (one by each ABR in the impacted area. If we generalize this to loss of reachability to N nodes throughout the network, the total number of additional advertisements is 2N. The received pulses can then be used to trigger BGP-PIC fast convergence.

The economy provided by the use of pulses in this use case diminishes linearly with the number of nodes which fail within a given time interval. In the event of a catastrophic network failure where many nodes fail within a given pulse interval, the number of pulses present in the network could begin to approach the number of individual prefixes present in the domain - which would effectively eliminate the scale benefits of the summary. Therefore, when using pulse for this use case, implementations SHOULD limit the number of pulses which are advertised in a given time interval.

#### 4.3. IS-IS Summary Component Reachability Loss Pulse TLV

IS-IS Summary Component Reachability Loss Pulse (SCRLP) TLV MAY be sent in an FSP-LSP. It is used by the IS-IS L1/L2 routers or by IS-IS Autonomous Boundary Routers (ASBR) that are performing prefix summarization at the area or domain boundary, to inform other nodes in the attached area(s) or domain(s) about the loss of the reachability to a previously reachable component of the summary-prefix inside the area or domain from which the summary-prefix is originated.

The IS-IS SCRLP TLV has the following format:

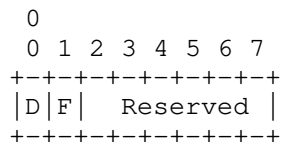


where:

Type: 1

Length: variable

Flags: 1 octet. The following flags are defined:



D-flag: Same as described in section 4.1. of [RFC5305].

F-flag: If unset, then the Summary Prefix and Component Prefix(es) are IPv4 prefixes. If set, then the Summary Prefix and Component Prefix(es) are IPv6 prefixes.

The remaining bits are reserved for future use. They MUST be set to zero on transmission and MUST be ignored on receipt.

R bits: reserved for future use. They MUST be set to zero on transmission and MUST be ignored on receipt.

MT ID: Multitopology Identifier as defined in [RFC5120]. Note that the value 0 is legal.

Summ-Pfx Length + Flag: 1 octet

Summ-Pfx Length: Length of the Summary Prefix in bits. Valid values are (0-31) when F-flag is unset, (0-127) when F-flag is set.

S-bit: MUST be set when Sub-TLVs are present for Summary Prefix, otherwise MUST NOT be set.

Summary Prefix: variable. IPv4 or IPv6 Summary Prefix.

Sub-TLV-length: 1 octet. Number of octets used by Summary Prefix Sub-TLVs. Only present when S-bit is set.

Optional Sub-TLVs: No Sub-TLVs are defined by this document.

Comp-Pfx Length + Flag: 1 octet

Comp-Pfx Len.: 1 octet. Length of the Component Prefix in bits. Valid values are (1-32) when F-flag is unset, (1-128) when F-flag is set. Comp-Pfx Len MUST be > Summ-Pfx Length.

S-bit: MUST be set when Sub-TLVs are present for Component Prefix, otherwise MUST NOT be set.

Component Prefix: variable. IPv4 or IPv6 Component Prefix.

Sub-TLV-length: 1 octet. Number of octets used by Component Prefix sub-TLVs. Only present when S-bit is set.

Optional sub-TLVs: No Sub-TLVs are defined by this document.

When an IS-IS L1/L2 router or an IS-IS Autonomous Boundary Router (ASBR) is performing prefix summarization and it loses the reachability to one or more previously reachable component(s) of the summary-prefix inside the area or domain for which the summarization is done, it MAY originate the SCRLP TLV to inform routers in other areas or domains about such summary component-prefix reachability loss.

An originator of the SCRLP TLV chooses to advertise it in FSP-LSP with L1 flooding scope and/or FSP-LSP with L2 flooding scope.

The IS-IS SCRLP TLV MAY be leaked between levels on L1/L2 router, subject to local policy of such L1/L2 router.

IS-IS SCRLP TLV MUST NOT be leaked inside the area if the summary prefix carried in IS-IS SCRLP TLV (Summary Prefix, Summ-Pfx Length) is advertised from such area by L1/L2 router.

When the router receives the SCRLP TLV it MAY choose to inform the BGP component on the router. BGP component on the router MAY trigger BGP Prefix Independent Convergence (PIC) as specified in [I-D.ietf-rtgwg-bgp-pic] as a result of such notification.

The mechanism on how the IS-IS passes the information from IS-IS SCRLP TLV to the BGP component or how the BGP component uses this information to trigger the PIC is implementation-specific and outside of the scope of this specification.

The IS-IS SCRLP TLV may be used by other applications on the receiving node that wish to be notified about the loss of summary component-prefix reachability. The details of such usage are outside of the scope of this specification.

## 5. Handling of the Control Plane Restart and ISSU

An egress PE may undergo a control-plane or protocol restart, or and In-Service Software Upgrade. If these events are performed using Nonstop Forwarding (NSF) as specified in [RFC3847] or Nonstop Routing (NSR) procedures, the egress PE reachability inside its area is preserved and no Pulse would be generated as a result of these events.

If the IS-IS protocol restart, or route-processor fail-over on the egress PE is performed using cold-restart procedures, the egress PE reachability will be lost and Pulse will be generated by the ABRs connected to the area. This is an expected behavior, as in case of the cold-restart recovery the traffic is expected to be dropped if forwarded to the egress PE and using an alternate BGP path is desirable.

## 6. OSPF Pulse Notification

TBD.

## 7. OSPFv3 Pulse Notification

TBD.

## 8. IANA Considerations

### 8.1. New IS-IS PDU Types

This document includes the definition of two new PDU types that are reflected in the "IS-IS PDU Registry":

Value	Description
7	FSP-LSP
8	FSP-PSNP

### 8.2. Revised sub-TLV table

IANA is requested to modify the table in "TLV Codepoints Registry" by adding columns for FSP-LSP and FSP-PSNP and set FSP-LSP:n and FSP-PSNP:n for all existing TLVs with the exception of 10 (Authentication) and 11 (ESN).

### 8.3. IS-IS Flooding Scope Pulse LSP Entries TLV

This document makes the following registrations in the IS-IS TLV Codepoints registry:

Type	Description	IIH	LSP	SNP	Purge	FSP-LSP	FSP-PSNP
29	FS-LSP Entries TLV	n	n	n	n	n	y

### 8.4. IS-IS Summary Component Reachability Loss Pulse TLV

This document makes the following registrations in the IS-IS TLV Codepoints registry:

Type	Description	IIH	LSP	SNP	Purge	FSP-LSP	FSP-PSNP
30	Summary Component Reachability Loss Pulse	n	n	n	n	y	n

## 9. Security Considerations for ISIS

The introduction of new PDU types introduces the possibility that an attacker could inject a false but apparently valid PDU. The use of

cryptographic authentication as defined in [RFC5304] or [RFC5310] minimizes the possibility of such occurrences.

Replay attacks could still be possible. Prevention of replay attacks can be done by including the Extended Sequence Numbers (ESN) TLV [RFC7602] in FSP-LSPs and FSP-PSNPs. Note, however, that the use of ESN MUST be done independently for each FSP-LSP ID. It is not safe to use a single ESN for the FSP-LSP PDU Type (as is done with hellos and SNPs) since we cannot guarantee the order in which multiple FSP-LSPs from the same source may arrive at a given node.

If a false PDU were to be injected, invalid SCRLP information could falsely trigger BGP-PIC behavior.

## 10. Security Considerations for OSPF

TBD.

## 11. Contributors

TBD

## 12. References

### 12.1. Normative References

- [ISO10589] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", 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>>.
- [RFC3847] Shand, M. and L. Ginsberg, "Restart Signaling for Intermediate System to Intermediate System (IS-IS)", RFC 3847, DOI 10.17487/RFC3847, July 2004, <<https://www.rfc-editor.org/info/rfc3847>>.
- [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>>.

- [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>>.
- [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>>.
- [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>>.
- [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>>.
- [RFC7602] Chunduri, U., Lu, W., Tian, A., and N. Shen, "IS-IS Extended Sequence Number TLV", RFC 7602, DOI 10.17487/RFC7602, July 2015, <<https://www.rfc-editor.org/info/rfc7602>>.
- [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>>.
- [RFC8706] Ginsberg, L. and P. Wells, "Restart Signaling for IS-IS", RFC 8706, DOI 10.17487/RFC8706, February 2020, <<https://www.rfc-editor.org/info/rfc8706>>.

## 12.2. Informative References

- [I-D.ietf-rtgwg-bgp-pic] Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", draft-ietf-rtgwg-bgp-pic-17 (work in progress), October 2021.

## Appendix A. BGP Pulse Handling

Handling of the Pulse by the receiving application is out of scope of this document. This section provides some informational, high-level description on how a BGP on an ingress PE (Provider Edge) device may use the Pulse to trigger the BGP PIC (Prefix Independent Convergence). Note that PIC is a local behavior on ingress PE, which is implementation specific and nothing in this section mandates the implementation to follow what is described here to any degree.

Assuming the BGP multi-path destination prefix on an ingress PE, the arrival of the Pulse, that indicates the loss of reachability of the BGP next-hop for the primary path, can trigger the BGP PIC for such prefix. This is similar in nature to what happens on ingress PE without the use of summarization when the BGP next-hop for primary path becomes unreachable.

In case the egress PE associated with the primary BGP path went down, BGP on the ingress PE would eventually receive a withdrawal of such path and would re-converge to the alternate path out of multi-paths. Note that this happens independently of and after the BGP PIC was triggered previously.

If the loss of reachability signaled by the Pulse is short-lived, it is desirable that BGP reconverge to the state prior to receipt of the Pulse. However, determination of the transient nature of the loss of reachability depends on the absence of BGP updates which would be expected following the loss of reachability to the egress PE. This can be determined by triggering a timer on receipt of the Pulse. If that timer expires without receipt of the expected BGP updates, then BGP can reconverge to the pre-pulse state. The timer duration needs to be long enough to allow for the expected BGP convergence to take place in the case where the loss of reachability to the egress PE is not transient.

#### Authors' Addresses

Peter Psenak (editor)  
Cisco Systems  
Pribinova Street 10  
Bratislava 81109  
Slovakia

Email: ppsenak@cisco.com

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

Email: ginsberg@cisco.com



Ketan Talaulikar  
Individual Contributor

Email: [ketant.ietf@gmail.com](mailto:ketant.ietf@gmail.com)

LSR Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2022

A. Wang  
China Telecom  
Z. Hu  
Huawei Technologies  
G. Mishra  
Verizon Inc.  
J. Sun  
ZTE Corporation  
July 12, 2021

Passive Interface Attribute  
draft-wang-lsr-passive-interface-attribute-08

Abstract

This document describes the mechanism that can be used to differentiate the passive interfaces from the normal interfaces within ISIS or OSPF domain.

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 13, 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. Conventions used in this document . . . . .	3
3. Consideration for flagging passive interface . . . . .	3
4. Passive Interface Attribute . . . . .	4
4.1. OSPFv2 Extended Stub-Link TLV . . . . .	4
4.2. OSPFv3 Router-Stub-Link TLV . . . . .	5
4.3. ISIS Stub-link TLV . . . . .	6
4.4. Stub-Link Prefix Sub-TLV . . . . .	7
5. Security Considerations . . . . .	8
6. IANA Considerations . . . . .	8
7. Acknowledgement . . . . .	9
8. References . . . . .	9
8.1. Normative References . . . . .	9
8.2. Informative References . . . . .	10
Authors' Addresses . . . . .	11

## 1. Introduction

Passive interfaces are used commonly within an operators enterprise or service provider networks. One of the most common use cases for passive interface is in a data center Layer 2 and Layer 3 Top of Rack(TOR) switch where the inter connected links between the TOR switches and uplinks to the Core switch are only a few links and a majority of the links are Layer 3 VLAN switched virtual interface trunked between the TOR switches serving Layer 2 broadcast domains. In this scenario all the VLANs are made passive as it is recommended to limit the number of network LSAs between routers and switches to avoid unnecessary hello processing overhead.

Another common use case is an inter-as routing scenario where the same routing protocol but different IGP instance is running between the adjacent BGP domains. Using passive interface on the inter-as connections can ensure that prefixes contained within a domain are only reachable within the domain itself and not allow the link state database to be merged between domain which could result in undesirable consequences.

For operator which runs different IGP domains that interconnect with each other via the passive interfaces, there is desire to obtain the inter-as topology information as described in [I-D.ietf-idr-bgpls-inter-as-topology-ext]. If the router that runs BGP-LS within one IGP domain can distinguish passive interfaces from

other normal interfaces, it is then easy for the router to report these passive links using BGP-LS to a centralized PCE controller.

Draft [I-D.dunbar-lsr-5g-edge-compute-ospf-ext] describes the case that edge compute server attach the network and needs to flood some performance index information to the network to facilitate the network select the optimized application resource. The edge compute server will also not run IGP protocol.

And, passive interfaces are normally the boundary of one IGP domain, knowing them can facilitate the operators to apply various policies on such interfaces, for example, to secure their networks, or filtering the incoming traffic with scrutiny.

But OSPF and ISIS have no position to flag such passive interface and their associated attributes now.

This document defines the protocol extension for OSPF and ISIS to indicate the passive interfaces and their associated attributes.

## 2. 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 [RFC2119] .

## 3. Consideration for flagging passive interface

ISIS [RFC5029] defines the Link-Attributes Sub-TLV to carry the link attribute information, but this Sub-TLV can only be carried within the TLV 22, which is used to described the attached neighbor. For passive interface, there is no ISIS neighbor, then it is not appropriate to use this Sub-TLV to indicate the passive attribute of the interface.

OSPFv2[RFC2328] defines link type field within Router LSA, the type 3 for connections to a stub network can be used to identified the passive interface. But in OSPFv3 [RFC5340], type 3 within the Router-LSA has been reserved. The information that associated with stub network has been put in the Intra-Area-Prefix-LSAs.

It is necessary to define one general solution for ISIS and OSPF to flag the passive interface and transfer the associated attributes then.

#### 4. Passive Interface Attribute

The following sections define the protocol extension to indicate the passive interface and associated attributes in OSPFv2/v3 and ISIS.

##### 4.1. OSPFv2 Extended Stub-Link TLV

[RFC7684] defines the OSPFv2 Extended Link Opaque LSA to contain the additional link attribute TLV. Currently, only OSPFv2 Extended Link TLV is defined to contain the link related sub-TLV. Because passive interface is not the normal link that participate in the OSPFv2 process, we select to define one new top TLV within the OSPFv2 Extended Link Opaque LSA to contain the passive interface related attribute information.

The OSPFv2 Extended Stub-Link TLV has the following format:

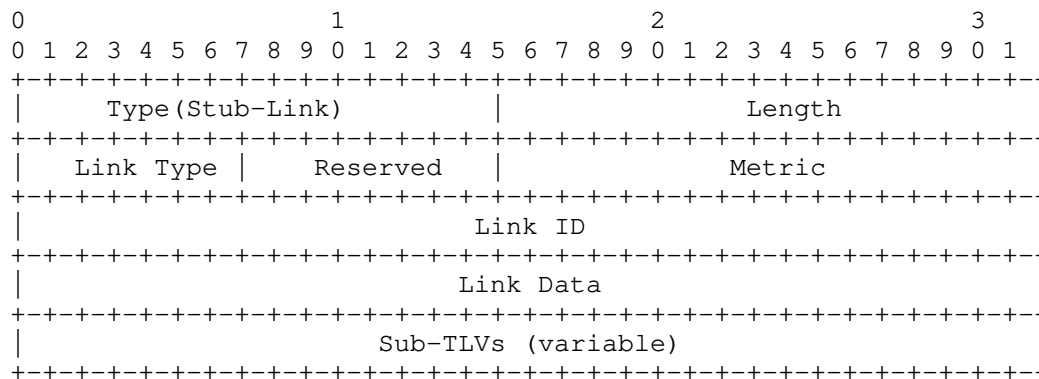


Figure 1: OSPFv2 Extended Stub-Link TLV

Type: The TLV type. The value is 2(TBD) for this stub-link type

Length: Variable, dependent on sub-TLVs

Link Type: Define the type of the stub-link. This document defines the followings type:

- o 0: Reserved
- o 1: AS boundary link
- o 2: Loopback link
- o 3: Vlan interface link
- o 4-255: For future extension

Metric: Link metric used for inter-AS traffic engineering.

Link ID: Link ID is defined in Section A.4.2 of [RFC2328]

Link Data: Link Data is defined in Section A.4.2 of [RFC2328]

Sub-TLVs: Existing sub-TLV that defined within "OSPFv2 Extended Link TLV Sub-TLV" can be included if necessary, the definition of new sub-TLV can refer to Section 4.4

If this TLV is advertised multiple times in the same OSPFv2 Extended Link Opaque LSA, only the first instance of the TLV is used by receiving OSPFv2 routers. This situation SHOULD be logged as an error.

If this TLV is advertised multiple times for the same link in different OSPFv2 Extended Link Opaque LSAs originated by the same OSPFv2 router, the OSPFv2 Extended Stub-Link TLV in the OSPFv2 Extended Link Opaque LSA with the smallest Opaque ID is used by receiving OSPFv2 routers. This situation may be logged as a warning.

It is RECOMMENDED that OSPFv2 routers advertising OSPFv2 Extended Stub-Link TLVs in different OSPFv2 Extended Link Opaque LSAs re-originate these LSAs in ascending order of Opaque ID to minimize the disruption.

This document creates a registry for Stub-Link attribute in Section 6.

#### 4.2. OSPFv3 Router-Stub-Link TLV

[RFC8362] extend the LSA format by encoding the existing OSPFv3 LSA [RFC5340] in TLV tuples and allowing advertisement of additional information with additional TLV.

This document defines the Router-Stub-Link TLV to describes a single router passive interface. The Router-Stub-Link TLV is only applicable to the E-Router-LSA. Inclusion in other Extended LSA MUST be ignored.

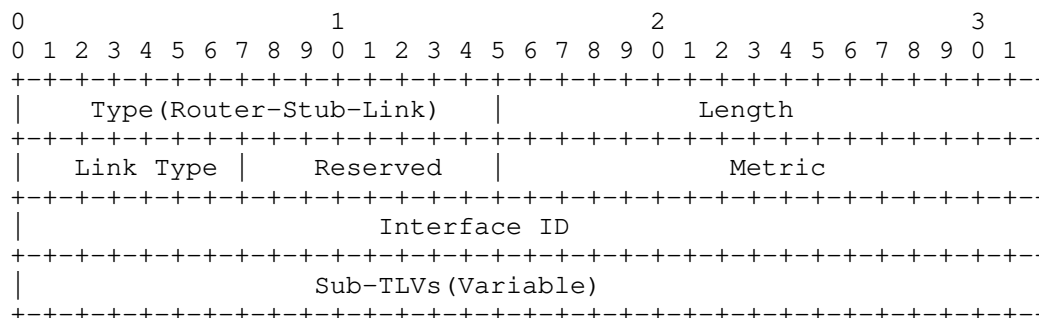


Figure 2: OSPFv3 Router-Stub-Link TLV

Type: OSPFv3 Extended-LSA TLV Type. Value is 10(TBD) for Router-Stub-Link TLV.

Length: Variable, dependent on sub-TLVs

Link Type: Define the type of the stub-link. This document defines the followings type:

- o 0: Reserved
- o 1: AS boundary link
- o 2: Loopback link
- o 3: Vlan interface link
- o 4-255: For future extension

Metric: Link metric used for inter-AS traffic engineering.

Interface ID: 32-bit number uniquely identifying this interface among the collection of this router's interfaces. For example, in some implementations it may be possible to use the MIB-II IfIndex [RFC2863].

Sub-TLVs: Existing sub-TLV that defined within "OSPFv3 Extended-LSA Sub-TLV" can be included if necessary. The definition of new sub-TLV can refer to Section 4.4.

#### 4.3. ISIS Stub-link TLV

This document defines one new top TLV to contain the passive interface attributes, which is shown in Figure 4:

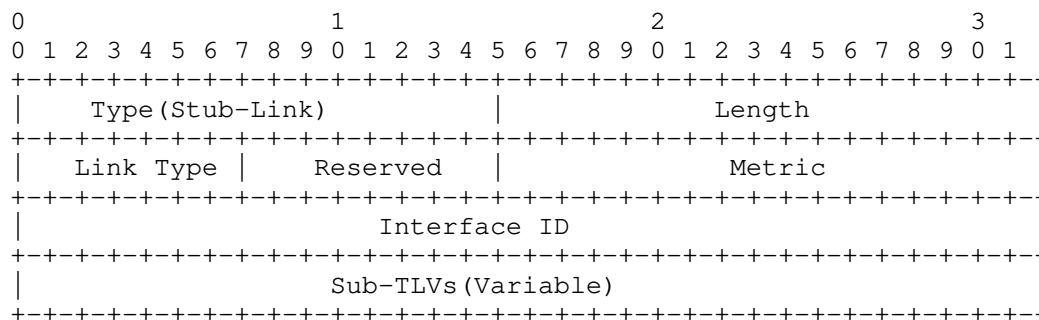


Figure 3: ISIS Stub-Link TLV

Type: ISIS TLV Codepoint. Value is 28(TBD) for stub-link TLV.

Length: Variable, dependent on sub-TLVs

Link Type: Define the type of the stub-link. This document defines the followings type:

- o 0: Reserved
- o 1: AS boundary link
- o 2: Loopback link
- o 3: Vlan interface link
- o 4-255: For future extension

Metric: Link metric used for inter-AS traffic engineering.

Interface ID: 32-bit number uniquely identifying this interface among the collection of this router's interfaces. For example, in some implementations it may be possible to use the MIB-II IfIndex [RFC2863].

Sub-TLVs: Existing sub-TLV that defined within "Sub-TLVs for TLVs 22, 23, 25, 141, 222, and 223" can be included if necessary. The definition of new sub-TLV can refer to Section 4.4.

#### 4.4. Stub-Link Prefix Sub-TLV

This document defines one new sub-TLV that can be contained within the OSPFv2 Extended Stub-Link TLV, OSPFv3 Router-Stub-Link TLV or ISIS Stub-Link TLV, to describe the prefix information associated with the passive interface.



The format of the sub-TLV is the followings:

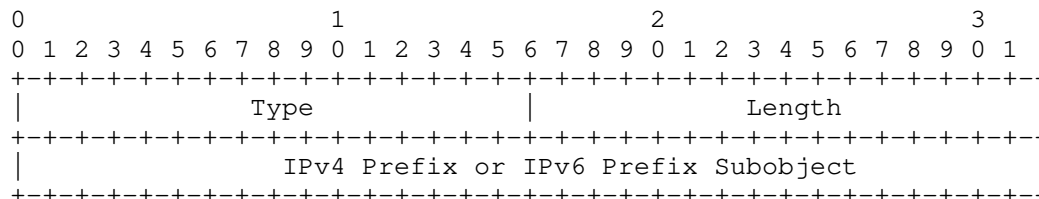


Figure 4: Stub-Link Prefix Sub-TLV

Type: The TLV type. The value is 01(TBD) for this Stub-Link Prefix type

Length: Variable, dependent on associated subobjects

Subobject: IPv4 prefix subobject or IPv6 prefix subobject, as that defined in [RFC3209]

If the passive interface has multiple address, then multiple subobjects will be included within this sub-TLV.

## 5. Security Considerations

Security concerns for ISIS are addressed in [RFC5304] and[RFC5310]

Security concern for OSPFv3 is addressed in [RFC4552]

Advertisement of the additional information defined in this document introduces no new security concerns.

## 6. IANA Considerations

IANA is requested to the allocation in following registries:

Registry	Type	Meaning
OSPFv2 Extended Link Opaque LSA TLV	2	Stub-Link TLV
OSPFv3 Extended-LSA TLV	10	Router-Stub-Link TLV
IS-IS TLV Codepoint	28	Stub-Link TLV

Figure 5: Newly defined TLV in existing IETF registry

IANA is requested to allocate one new registry that can be referred by OSPFv2, OSPFv3 and ISIS respectively.

New Registry	Meaning
Stub-Link Attribute	Attributes for stub-link

Figure 6: Newly defined Registry for stub-link attributes

One new sub-TLV is defined in this document under this registry codepoint:

Registry	Type	Meaning
Stub-Link Attribute	0	Reserved
	1	Stub-Link Prefix sub-TLV
	2-65535	Reserved

Figure 7: Stub-Link Prefix Sub-TLV

## 7. Acknowledgement

Thanks Shunwan Zhang, Tony Li, Les Ginsberg, Acee Lindem, Dhruv Dhody, Jeff Tantsura and Robert Raszuk for their suggestions and comments on this idea.

## 8. References

### 8.1. 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, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [RFC2863] McCloghrie, K. and F. Kastenholtz, "The Interfaces Group MIB", RFC 2863, DOI 10.17487/RFC2863, June 2000, <<https://www.rfc-editor.org/info/rfc2863>>.

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC4552] Gupta, M. and N. Melam, "Authentication/Confidentiality for OSPFv3", RFC 4552, DOI 10.17487/RFC4552, June 2006, <<https://www.rfc-editor.org/info/rfc4552>>.
- [RFC5029] Vasseur, JP. and S. Previdi, "Definition of an IS-IS Link Attribute Sub-TLV", RFC 5029, DOI 10.17487/RFC5029, September 2007, <<https://www.rfc-editor.org/info/rfc5029>>.
- [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>>.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <<https://www.rfc-editor.org/info/rfc5340>>.
- [RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <<https://www.rfc-editor.org/info/rfc7684>>.
- [RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <<https://www.rfc-editor.org/info/rfc7794>>.
- [RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <<https://www.rfc-editor.org/info/rfc8362>>.

## 8.2. Informative References

- [I-D.dunbar-lsr-5g-edge-compute-ospf-ext]  
Dunbar, L., Chen, H., and A. Wang, "OSPF extension for 5G Edge Computing Service", draft-dunbar-lsr-5g-edge-compute-ospf-ext-04 (work in progress), March 2021.

[I-D.ietf-idr-bgppls-inter-as-topology-ext]

Wang, A., Chen, H., Talaulikar, K., and S. Zhuang, "BGP-LS  
Extension for Inter-AS Topology Retrieval", draft-ietf-  
idr-bgppls-inter-as-topology-ext-09 (work in progress),  
September 2020.

#### Authors' Addresses

Aijun Wang  
China Telecom  
Beiqijia Town, Changping District  
Beijing 102209  
China

Email: wangaj3@chinatelecom.cn

Zhibo Hu  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: huzhibo@huawei.com

Gyan S. Mishra  
Verizon Inc.  
13101 Columbia Pike  
Silver Spring MD 20904  
United States of America

Email: gyan.s.mishra@verizon.com

Jinsong Sun  
ZTE Corporation  
No. 68, Ziiijnhua Road  
Nan Jing 210012  
China

Email: sun.jinsong@zte.com.cn

LSR Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 18, 2022

A. Wang  
China Telecom  
G. Mishra  
Verizon Inc.  
Z. Hu  
Y. Xiao  
Huawei Technologies  
October 15, 2021

Prefix Unreachable Announcement  
draft-wang-lsr-prefix-unreachable-announcement-08

Abstract

This document describes a mechanism to solve an existing issue with Longest Prefix Match (LPM), that exists where an operator domain is divided into multiple areas or levels where summarization is utilized. This draft addresses a fail-over issue related to a multi areas or levels domain, where a link or node down event occurs resulting in an LPM component prefix being omitted from the FIB resulting in black hole sink of routing and connectivity loss. This draft introduces a new control plane convergence signaling mechanism using a negative prefix called Prefix Unreachable Announcement Mechanism(PUAM), utilized to detect a link or node down event and signal the RIB that the event has occurred to force immediate control plane convergence.

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 18, 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. Conventions used in this document . . . . .	3
3. Scenario Description . . . . .	3
3.1. Inter-Area Node Failure Scenario . . . . .	4
3.2. Inter-Area Links Failure Scenario . . . . .	4
4. PUA (Prefix Unreachable Advertisement) Procedures . . . . .	5
5. MPLS and SRv6 LPM based BGP Next-hop Failure Application . . . . .	5
6. PUAM Capabilities Announcement . . . . .	6
7. Implementation Consideration . . . . .	7
8. Deployment Considerations . . . . .	7
9. Security Considerations . . . . .	8
10. IANA Considerations . . . . .	8
11. Acknowledgement . . . . .	9
12. Normative References . . . . .	9
Authors' Addresses . . . . .	10

## 1. Introduction

As part of an operator optimized design criteria, a critical requirement is to limit Shortest Path First (SPF) churn which occurs within a single OSPF area or ISIS level. This is accomplished by sub-dividing the IGP domain into multiple areas for flood reduction of intra area prefixes so they are contained within each discrete area to avoid domain wide flooding.

OSPF and ISIS have a default and summary route mechanism which is performed on the OSPF area border router or ISIS L1-L2 node. The OSPF summary route is triggered to be advertised conditionally when at least one component prefix exists within the non-zero area. ISIS Level-L1-L2 node as well generate a summary prefix into the level-2 backbone area for Level 1 area prefixes that is triggered to be

advertised conditionally when at least a single component prefix exists within the Level-1 area. ISIS L1-L2 node with attach bit set also generates a default route into each Level-1 area along with summary prefixes generated for other Level-1 areas.

Operators have historically relied on MPLS architecture which is based on exact match host route FEC binding for single area. [RFC5283] LDP inter-area extension provides the ability to LPM, so now the RIB match can now be a summary match and not an exact match of a host route of the egress PE for an inter-area LSP to be instantiated. SRV6 routing framework utilizes the IPv6 data plane standard IGP LPM. When operators start to migrate from MPLS LSP based host route bootstrapped FEC binding, to SRV6 routing framework, the IGP LPM now comes into play with summarization which will influence the forwarding of traffic when a link or node event occurs for a component prefix within the summary range resulting in black hole routing of traffic.

The motivation behind this draft is based on either MPLS LPM FEC binding, or SRv6 BGP service overlay using traditional unicast routing (uRIB) LPM forwarding plane where the IGP domain has been carved up into OSPF or ISIS areas and summarization is utilized. In this scenario where a failure conditions result in a black hole of traffic where multiple ABRs exist and either the area is partitioned or other link or node failures occur resulting in the component prefix host route missing within the summary range. Summarization of inter-area types routes propagated into the backbone area for flood reduction are made up of component prefixes. It is these component prefixes that the PUAM tracks to ensure traffic is not black hole sink routed due to a PE or ABR failure. The PUA mechanism ensures immediate control plane convergence with ABR or PE node switchover when area is partitioned or ABR has services down to avoid black hole of traffic.

## 2. 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 [RFC2119] .

## 3. Scenario Description

Figure 1 illustrates the topology scenario when OSPF or ISIS is running in multi areas or multi levels domain. R0-R4 are routers in backbone area, S1-S4,T1-T4 are internal routers in area 1 and area 2 respectively. R1 and R3 are area border routers or ISIS Level 1-2 border nodes between area 0 and area 1. R2 and R4 are area border routers between area 0 and area 2.

S1/S4 and T2/T4 PEs peer to customer CEs for overlay VPNs. Ps1/Ps4 is the loopback0 address of S1/S4 and Pt2/Pt4 is the loopback0 address of T2/T4.

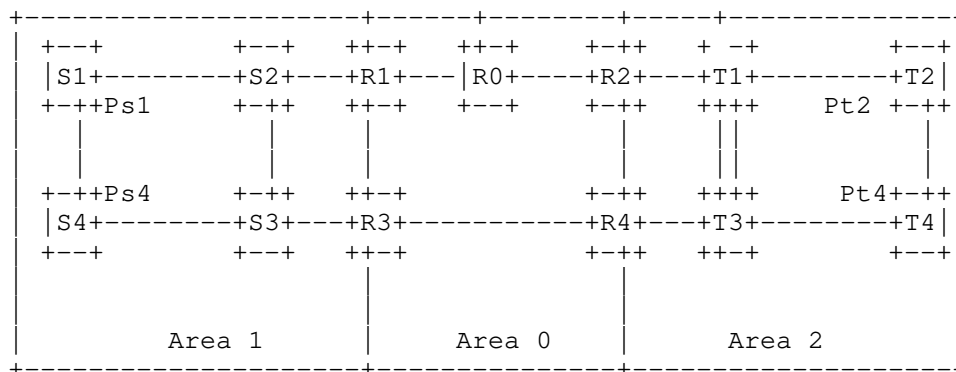


Figure 1: OSPF Inter-Area Prefix Unreachable Announcement Scenario

### 3.1. Inter-Area Node Failure Scenario

If the area border router R2/R4 does the summary action, then one summary address that cover the prefixes of area 2 will be announced to area 0 and area 1, instead of the detail address. When the node T2 is down, Pt2 bgp next hop becomes unreachable while the LPM summary prefix continues to be advertised into the backbone area. Except the border router R2/R4, the other routers within area 0 and area 1 do not know the unreachable status of the Pt2 bgp next hop prefix. Traffic will continue to forward LPM match to prefix Pt2 and will be dropped on the ABR or Level 1-2 border node resulting in black hole routing and connectivity loss. Customer overlay VPN dual homed to both S1/S4 and T2/R4, traffic will not be able to fail-over to alternate egress PE T4 bgp next hop Pt4 due to the summarization.

### 3.2. Inter-Area Links Failure Scenario

In a link failure scenario, if the link between T1/T2 and T1/T3 are down, R2 will not be able to reach node T2. But as R2 and R4 do the summary announcement, and the summary address covers the bgp next hop prefix of Pt2, other nodes in area 0 area 1 will still send traffic to T2 bgp next hop prefix Pt2 via the border router R2, thus black hole sink routing the traffic.

In such a situation, the border router R2 should notify other routers that it can't reach the prefix Pt2, and lets the other ABRs(R4) that can reach prefix Pt2 advertise one specific route to Pt2, then the



internal routers will select R4 as the bypass router to reach prefix Pt2.

#### 4. PUA (Prefix Unreachable Advertisement) Procedures

[RFC7794] and [I-D.ietf-lsr-ospf-prefix-originator] draft both define one sub-tlv to announce the originator information of the one prefix from a specified node. This draft utilizes such TLV for both OSPF and ISIS to signal the negative prefix in the perspective PUAM when a link or node goes down.

ABR detects link or node down and floods PUAM negative prefix advertisement along with the summary advertisement according to the prefix-originator specification. The ABR or ISIS L1-L2 border node has the responsibility to add the prefix originator information when it receives the Router LSA from other routers in the same area or level.

When the ABR or ISIS L1-L2 border node generates the summary advertisement based on component prefixes, the ABR will announce one new summary LSA or LSP which includes the information about this down prefix, with the prefix originator set to NULL. The number of PUAMs is equivalent to the number of links down or nodes down. The LSA or LSP will be propagated with standard flooding procedures.

If the nodes in the area receive the PUAM flood from all of its ABR routers, they will start BGP convergence process if there exist BGP session on this PUAM prefix. The PUAM creates a forced fail over action to initiate immediate control plane convergence switchover to alternate egress PE. Without the PUAM forced convergence the down prefix will yield black hole routing resulting in loss of connectivity.

When only some of the ABRs can't reach the failure node/link, as that described in Section 3.2, the ABR that can reach the PUAM prefix should advertise one specific route to this PUAM prefix. The internal routers within another area can then bypass the ABRs that can't reach the PUAM prefix, to reach the PUAM prefix.

#### 5. MPLS and SRv6 LPM based BGP Next-hop Failure Application

In an MPLS or SR-MPLS service provider core, scalability has been a concern for operators which have split up the IGP domain into multiple areas to avoid SPF churn. Normally, MPLS FEC binding for LSP instantiation is based on egress PE exact match of a host route Looback0. [RFC5283] LDP inter-area extension provides the ability to LPM, so now the RIB match can now be a summary match and not an exact match of host route of the egress PE for an inter-area LSP to be

instantiated. The caveat related to this feature that has prevented operators from using the [RFC5283] LDP inter-area extension concept is that when the component prefixes are now hidden in the summary prefix, and thus the visibility of the BGP next-hop attribute is lost.

In a case where a PE is down, and the [RFC5283] LDP inter-area extension LPM summary is used to build the LSP inter-area, the LSP remains partially established black hole on the ABR performing the summarization. This major gap with [RFC5283] inter-area extension forces operators into a workaround of having to flood the BGP next-hop domain wide. In a small network this is fine, however if you have 1000s PEs and many areas, the domain wide flooding can be painful for operators as far as resource usage memory consumption and computational requirements for RIB / FIB / LFIB label binding control plane state. The ramifications of domain wide flooding of host routes is described in detail in [RFC5302] domain wide prefix distribution with 2 level ISIS Section 1.2 - Scalability. As SRv6 utilizes LPM, this problem exists as well with SRv6 when IGP domain is broken up into areas and summarization is utilized.

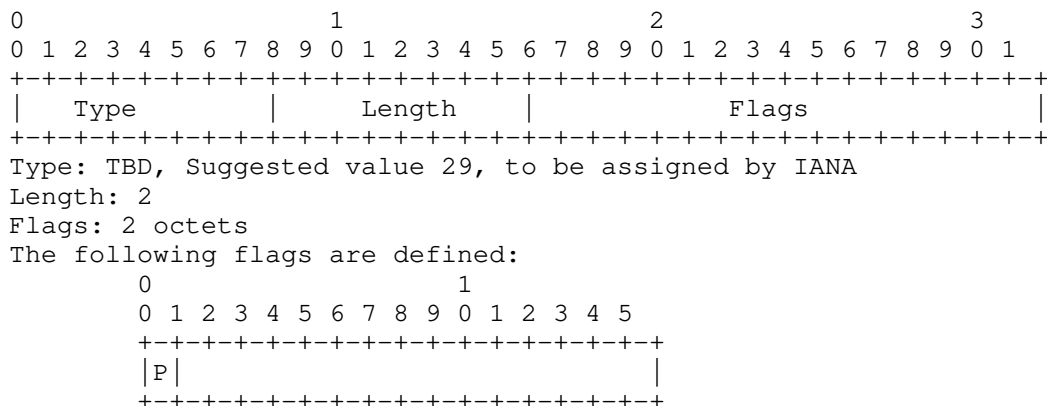
PUAM is now able to provide the negative prefix component flooded across the backbone to the other areas along with the summary prefix, which is now immediately programmed into the RIB control plane. MPLS LSP exact match or SRv6 LPM match over fail over path can now be established to the alternate egress PE. No disruption in traffic or loss of connectivity results from PUAM. Further optimizations such as LFA and BFD can be done to make the data plane convergence hitless. The PUAM solution applies to MPLS or SR-MPLS where LDP inter-area extension is utilized for LPM aggregate FEC, as well a SRv6 IPv6 control plane LPM match summarization of BGP next hop.

## 6. PUAM Capabilities Announcement

When not all of the nodes in one area support the PUAM information, there are possibilities to form traffic loop. To avoid this happen, the ABR should not send PUAM information to one area until it ensures that all of nodes in this area can parse the PUAM information. To accomplish this, this draft defines the capabilities sub-TLV as the followings:

For OSPFv2, this bit (Bit number TBD, suggest bit 6, 0x20) should be carried in "OSPF Router-LSA Option", as that described in [RFC2328]. For OSPFv3, one bit (Bit number TBD, suggest bit 8) should be defined to indicate the router's capabilities to support PUAM that described in this draft, the defined bit should be carried in "OSPF Router Informational Capabilities" TLV, which is described in [RFC7770]. For ISIS, one new sub-TLV(Type TBD, suggest 29), PUAM Capabilities

sub-TLV, which is included in the "IS-IS Router CAPABILITY TLV" [RFC7981] is defined in the followings:



where:

P-flag: If set, the router supports PUA information.

Figure 2: PUA Capabilities sub-TLV format

## 7. Implementation Consideration

Considering the balances of reachable information and unreachable information announcement capabilities, the implementation of this mechanism should set one MAX\_Address\_Announcement (MAA) threshold value that can be configurable. Then, the ABR should make the following decisions to announce the prefixes:

1. If the number of unreachable prefixes is less than MAA, the ABR should advertise the summary address and the PUAM.
2. If the number of reachable address is less than MAA, the ABR should advertise the detail reachable address only.
3. If the number of reachable prefixes and unreachable prefixes exceed MAA, then advertise the summary address with MAX metric.

## 8. Deployment Considerations

To support the PUAM advertisement, the ABRs should be upgraded according to the procedures described in Section 4. The PEs that want to accomplish the BGP switchover that described in Section 3.1 and Section 5 should also be upgraded to act upon the receive of the PUAM message. Other nodes within the network can ignore such PUAM message if they don't care or don't support.

As described in Section 4, the ABR will advertise the PUAM message once it detects there is link or node down within the summary address. In order to reduce the unnecessary advertisements of PUAM messages on ABRs, the ABRs should support the configuration of the protected prefixes. Based on such information, the ABR will only advertise the PUAM message when the protected prefixes (for example, the loopback addresses of PEs that run BGP) that within the summary address is missing.

The advertisement of PUAM message should only last one configurable period to allow the services that run on the failure prefixes are converged or switchover. If one prefix is missed before the PUAM takes effect, the ABR will not declare its absence via the PUAM.

## 9. Security Considerations

Advertisement of PUAM information follow the same procedure of traditional LSA. The action based on the PUAM is clearly defined in this document for ABR or Level1/2 router and the receiver that run BGP.

There is no changes to the forward behavior of other internal routers.

## 10. IANA Considerations

IANA is requested to register the following in the "OSPF Router Properties Registry" and "OSPF Router Informational Capability Bits Registry" respectively.

Bit Number	Capability Name	Reference
TBD(0x20)	OSPF PUA Support	this document

Table 1: P-Bit in OSPF Router-LSA Option

Bit Number	Capability Name	Reference
TBD(bit 8)	OSPF PUA Support	this document

Table 2: OSPF Router PUA Capability Support Bit

IANA is requested to register the following in "Sub-TLVs for TLV242 (IS-IS Router CAPABILITY TLV)

Type: 29 (Suggested - to be assigned by IANA)

Description: PUA Support Capabilities

## 11. Acknowledgement

Thanks Peter Psenak, Les Ginsberg, Acee Lindem, Shraddha Hegde, Robert Raszuk, Tonly Li, Jeff Tantsura, Tony Przygienda and Bruno Decraene for their suggestions and comments on this draft.

## 12. Normative References

- [I-D.ietf-lsr-ospf-prefix-originator]  
Wang, A., Lindem, A., Dong, J., Psenak, P., and K. Talaulikar, "OSPF Prefix Originator Extensions", draft-ietf-lsr-ospf-prefix-originator-12 (work in progress), April 2021.
- [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>>.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [RFC5283] Decraene, B., Le Roux, J.L., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, DOI 10.17487/RFC5283, July 2008, <<https://www.rfc-editor.org/info/rfc5283>>.

- [RFC5302] Li, T., Smit, H., and T. Przygienda, "Domain-Wide Prefix Distribution with Two-Level IS-IS", RFC 5302, DOI 10.17487/RFC5302, October 2008, <<https://www.rfc-editor.org/info/rfc5302>>.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <<https://www.rfc-editor.org/info/rfc5340>>.
- [RFC5709] Bhatia, M., Manral, V., Fanto, M., White, R., Barnes, M., Li, T., and R. Atkinson, "OSPFv2 HMAC-SHA Cryptographic Authentication", RFC 5709, DOI 10.17487/RFC5709, October 2009, <<https://www.rfc-editor.org/info/rfc5709>>.
- [RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <<https://www.rfc-editor.org/info/rfc7770>>.
- [RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <<https://www.rfc-editor.org/info/rfc7794>>.
- [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>>.

## Authors' Addresses

Aijun Wang  
China Telecom  
Beiqijia Town, Changping District  
Beijing 102209  
China

Email: wangaj3@chinatelecom.cn

Gyan Mishra  
Verizon Inc.

Email: gyan.s.mishra@verizon.com

Zhibo Hu  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: [huzhibo@huawei.com](mailto:huzhibo@huawei.com)

Yaqun Xiao  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: [xiaoyaqun@huawei.com](mailto:xiaoyaqun@huawei.com)