

LSR
Internet-Draft
Intended status: Standards Track
Expires: 8 May 2022

R. Chen
D. Zhao
ZTE Corporation
P. Psenak
K. Talaulikar
Cisco Systems
4 November 2021

Updates to Anycast Property advertisement for OSPF
draft-chen-lsr-anycast-flag-01

Abstract

Each prefix is advertised along with an 8-bit field of capabilities, by using the Prefix Options [RFC8362] and the flag field in the OSPFv2 Extended Prefix TLV [RFC7684], but the definition of anycast flag to identify the prefix as anycast has not yet been defined. However, Almost all bits of the Flag field has been assigned already. Thus, it is also required to extend the flag field for future use.

This document updates [RFC7684] and [RFC8362], by defining a new variable length Prefix attributes Sub-TLVs for OSPFv2 and OSPFv3 and a new flag in the Prefix attributes Sub-TLV to advertise the anycast property.

Status of This Memo

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

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

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

This Internet-Draft will expire on 8 May 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	2
1.1. Requirements Language	3
2. Variable length Prefix attributes Sub-TLV	3
3. Processing	5
4. Acknowledgements	6
5. IANA Considerations	6
5.1. OSPFv2 Extended Prefix Sub-TLV Registry	6
5.2. OSPFv3 Extended LSA Sub-TLV Registry	6
6. Security Considerations	7
7. Normative References	7
Authors' Addresses	7

1. Introduction

Both SR-MPLS prefixes-SID and IPv4/IPv6 prefix may be configured as anycast and as such the same value can be advertised by multiple routers. It is useful for other routers to know that the advertisement is for an anycast identifier.

[RFC7684] defines OSPFv2 Opaque LSAs based on Type-Length-Value (TLV) tuples that can be used to associate additional attributes with prefixes or links. 8-bit field of the OSPFv2 Extended Prefix TLV is used to advertise additional attributes associated with the prefix, but the definition of anycast flag to identify the prefix as anycast has not yet been defined. However, three bits have been defined.

[RFC8362] extends the LSA format by encoding the existing OSPFv3 LSA information in Type-Length-Value (TLV) tuples and allowing advertisement of additional information with additional TLVs. Each prefix is advertised along with an 8-bit field of capabilities, by using the Prefix Options, but the definition of anycast flag to identify the prefix as anycast has not yet been defined. However, only the final bit in the Prefix Options is not allocated.

This document updates [RFC7684] and [RFC8362], by defining a new variable length Prefix attributes Sub-TLVs for OSPFv2 and OSPFv3 and a new flag in the Prefix attributes Sub-TLV to advertise the anycast property.

1.1. 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].

2. Variable length Prefix attributes Sub-TLV

This document creates a new variable length Prefix attributes Sub-TLV for OSPFv2 and OSPFv3. This Sub-TLV specifies a variable flag fields to advertise additional attributes associated with the prefix.

The format of each TLV is:

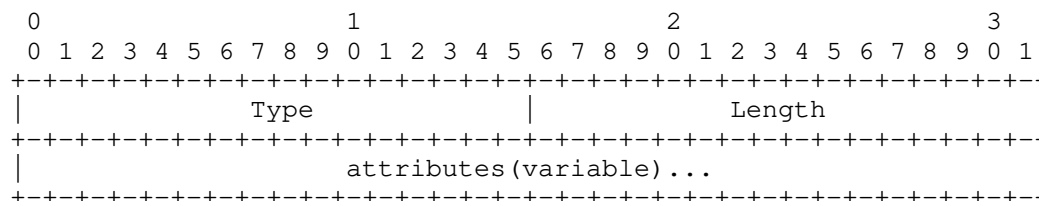


Figure 1

where:

Type: TBA.

Length: Variable, dependent on the included attributes.

Attributes: Variable. The extended flag fields, and the first 8 bits are reserved for the flag field previously defined by OSPFv2 and OSPFv3.

In the case of OSPFv2, the Prefix attributes Sub-TLVs is a sub-TLV of the OSPFv2 Extended Prefix TLV as defined in [RFC7684]. Figure 2 below is the definition of attribute field.

Attributes: The following flags are defined and the first 8 bits are reserved for the previously defined one-octet field contains flags in OSPFv2 Extended Prefix TLV [RFC7684]:

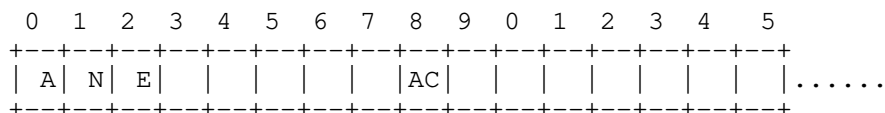


Figure 2

Where:

E-Flag: Refer to [RFC9089] .

N/A-Flag: Refer to [RFC7684].

AC-flag: A new flag is used to advertise the anycast property. When the prefix is configured as anycast, the AC-flag SHOULD be set. Otherwise, this flag MUST be clear. If both N-flag and AC-flag are set, the receiving routers MUST ignore the N-flag.

AC-flag MUST be preserved when the prefix is propagated between areas.

The same prefix can be advertised by multiple routers, and that if at least one of them sets the AC-Flag in its advertisement, the prefix SHOULD be considered as anycast.

The other bits are reserved for future use.

In the case of OSPFv3, the Prefix attributes Sub-TLVs is a sub-TLV of the following OSPFv3 TLVs as defined in [RFC8362]:

- * Intra-Area Prefix TLV
- * Inter-Area Prefix TLV
- * External Prefix TLV

Figure 3 below is the definition of attribute field.

Attributes: The following flags are defined and the first 8 bits are reserved for the previously defined OSPFv3 Prefix Options:

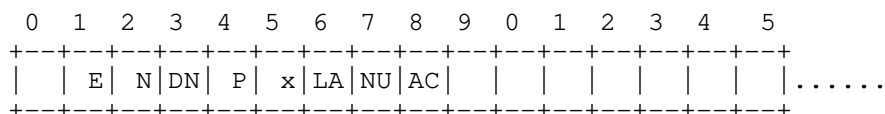


Figure 3

Where:

E-Flag: Refer to [RFC9089] in section 3.1.

N-Flag: Refer to [RFC8362] in section 3.1.1.

P/x/LA/NU-Flag: Refer to [RFC5340].

AC-Flag: A new flag is used to advertise the anycast property. When the prefix is configured as anycast, the AC-flag SHOULD be set. Otherwise, this flag MUST be clear. If both N-flag and AC-flag are set, the receiving routers MUST ignore the N-flag.

AC-flag MUST be preserved when the prefix is propagated between areas.

The same prefix can be advertised by multiple routers, and that if at least one of them sets the AC-Flag in its advertisement, the prefix SHOULD be considered as anycast.

The other bits are reserved for future use.

3. Processing

If there is an device in the network that does not support the Extension of the Prefix attributes Sub-TLV, then the device that support the Extension of the Prefix attributes Sub-TLV should advertise the field of capabilities of the Prefix by using prefix-options[RFC8362] or prefix-flags[RFC7684], and the Prefix attributes Sub-TLV. Otherwise, only use the Prefix attributes Sub-TLV to advertise the field of capabilities of the Prefix.

If prefix is advertised along with the field of capabilities, by using the Prefix attributes Sub-TLV, then the field of capabilities of the Prefix in the OSPFv2/OSPFv3 Prefix attributes Sub-TLV shall prevail.

As long as the Prefix attributes Sub-TLV is used to advertise the field of capabilities and the device support the Extension of the Prefix attributes Sub-TLV, then the field of capabilities in the Prefix attributes Sub-TLV shall prevail.

If prefix is advertised along with the field of capabilities, by using only the prefix-options[RFC8362] or prefix-flags[RFC7684], then the field of capabilities in the prefix-options[RFC8362] or prefix-flags[RFC7684] shall prevail.

4. Acknowledgements

TBD.

5. IANA Considerations

This document requests allocation for the following registry.

5.1. OSPFv2 Extended Prefix Sub-TLV Registry

This document requests allocation for OSPFv2 Extended Prefix Sub-TLV Registry:

Value	Description	Reference
TBA	OSPFv2 Prefix attributes Sub-TLV	This document

Figure 4

This document adds a new bit in the "OSPFv2 Prefix attributes Sub-TLV" registry:

AC-flag (Anycast Flag).

5.2. OSPFv3 Extended LSA Sub-TLV Registry

This document requests allocation for OSPFv3 Extended LSA Sub-TLV Registry:

Value	Description	Reference
TBA	OSPFv3 Prefix attributes Sub-TLV	This document

Figure 5

This document adds a new bit in the "OSPFv3 Prefix attributes TLV" registry:

AC-flag (Anycast Flag).

6. Security Considerations

Procedures and protocol extensions defined in this document do not affect the OSPFv2 , OSPFv3 security model. See the "Security Considerations" section of [RFC7684] for a discussion of OSPFv2 security, the "Security Considerations" section of [RFC8362] for a discussion of OSPFv3 security.

7. 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>>.
- [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>>.
- [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>>.
- [RFC9089] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using OSPF", RFC 9089, DOI 10.17487/RFC9089, August 2021, <<https://www.rfc-editor.org/info/rfc9089>>.

Authors' Addresses

Ran Chen
ZTE Corporation
Nanjing
China

Email: chen.ran@zte.com.cn

Detao Zhao
ZTE Corporation
Nanjing
China

Email: zhao.detao@zte.com.cn

Peter Psenak
Cisco Systems
Slovakia

Email: ppsenak@cisco.com

Ketan Talaulikar
Cisco Systems
India

Email: ketant@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 25 April 2022

B. Decraene
Orange
L. Ginsberg
Cisco Systems
T. Li
Arista Networks
G. Solignac

M. Karasek
Cisco Systems
C. Bowers
Juniper Networks, Inc.
G. Van de Velde
Nokia
P. Psenak
Cisco Systems
T. Przygienda
Juniper
22 October 2021

IS-IS Fast Flooding
draft-decraeneginsberg-lsr-isis-fast-flooding-00

Abstract

Current Link State Protocol Data Unit (PDU) flooding rates 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 the need for faster flooding, the issues around faster flooding, and some example approaches to achieve faster flooding. It also defines protocol extensions relevant to faster flooding.

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 25 April 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
2. Requirements Language	3
3. Historical Behavior	4
4. Flooding Parameters TLV	5
4.1. LSP Burst Window sub-TLV	6
4.2. LSP Transmission Interval sub-TLV	6
4.3. LSPs Per PSNP sub-TLV	6
4.4. Flags sub-TLV	6
4.5. Partial SNP Interval sub-TLV	7
4.6. Operation on a LAN interface	7
5. Performance improvement on the receiver	8
5.1. Rate of LSP Acknowledgments	8
5.2. Packet Prioritization on Receive	9
6. Congestion and Flow Control	10
6.1. Overview	10
6.2. Congestion and Flow Control algorithm: Example 1	10
6.3. Congestion Control algorithm: Example 2	17
7. IANA Considerations	19
8. Security Considerations	20
9. Contributors	21
10. Acknowledgments	21
11. References	21
11.1. Normative References	21
11.2. Informative References	22
Appendix A. Changes / Author Notes	22
Appendix B. Issues for Further Discussion	22
Authors' Addresses	22

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 is the result of guidance in the base specification [ISO10589] and early deployments when both CPU speeds and interface speeds were much slower and the scale of an area was much smaller than they are today.

As IS-IS is deployed in greater scale both in the number of nodes in an area and in the number of neighbors per 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 typical LSP flooding rates used 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 the LSP flooding rate therefore becomes an essential element of supporting greater network scale.

Improving the LSP flooding rate is complementary to protocol extensions that reduce LSP flooding traffic by reducing the flooding topology such as Mesh Groups [RFC2973] or Dynamic Flooding [I-D.ietf-lsr-dynamic-flooding]. Reduction of the flooding topology does not alter the number of LSPs required to be exchanged between two nodes, so increasing the overall flooding speed is still beneficial when such extensions are in use. It is also possible that the flooding topology can be reduced in ways that prefer the use of neighbors that support improved flooding performance.

2. 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.

3. 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 is 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 is 5 seconds.

`partialSNPInterval` - This is the amount of time between periodic action for transmission of Partial Sequence Number PDUs. It shall be less than `minimumLSPTransmissionInterval`.

The recommended value is 2 seconds.

Most relevant to a discussion of the 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. It is 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 the 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, the flooding rate has not been fundamentally altered.

4. Flooding Parameters TLV

This document defines a new Type-Length-Value tuple (TLV) called the "Flooding Parameters TLV" that may be included in IS to IS Hellos (IIH) or Partial Sequence Number PDUs (PSNPs). It allows IS-IS implementations to advertise flooding related parameters and capabilities which may be of use to the peer in support of faster flooding.

Type: TBD1

Length: variable, the size in octets of the Value field

Value: One or more sub-TLVs

Several sub-TLVs are defined in this document. The support of any sub-TLV is OPTIONAL.

For a given IS-IS adjacency, the Flooding Parameters TLV does not need to be advertised in each IIH or PSNP. An IS uses the latest received value for each parameter until a new value is advertised by the peer. However, as IIHs and PSNPs are not reliably exchanged, and may never be received, parameters SHOULD be sent even if there is no change in value since the last transmission. For a parameter which has never been advertised, an IS SHOULD use its local default value. That value SHOULD be configurable on a per node basis and MAY be configurable on a per interface basis.

4.1. LSP Burst Window sub-TLV

The LSP Burst Window sub-TLV advertises the maximum number of LSPs that the node can receive with no separation interval between LSPs.

Type: 1

Length: 4 octets

Value: number of LSPs that can be sent back to back

4.2. LSP Transmission Interval sub-TLV

The LSP Transmission Interval sub-TLV advertises the minimum interval, in micro-seconds, between LSPs arrivals which can be 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 burst window has been used

The LSP Transmission Interval is an advertisement of the receiver's steady-state LSP reception rate.

4.3. LSPs Per PSNP sub-TLV

The LSP per PSNP (LPP) sub-TLV advertises the number of received LSPs that triggers the immediate sending of a PSNP to acknowledge them.

Type: 3

Length: 2 octets

Value: number of LSPs acknowledged per PSNP

A node advertising this sub-TLV with a value LPP MUST send a PSNP once LPP LSPs have been received and need to be acknowledged.

4.4. Flags sub-TLV

The sub-TLV Flags advertises a set of flags.

Type: 4

Length: Indicates the length in octets (1-8) of the Value field. The length SHOULD be the minimum required to send all bits that are set.

Value: List of flags.

```

    0 1 2 3 4 5 6 7 ...
    +-+-+-+-+-+-+-+...
    |0|               ...
    +-+-+-+-+-+-+-+...
```

When the 0 flag is set, the LSP will be acknowledged in the order they are received: a PSNP acknowledging N LSPs is acknowledging the N oldest LSPs received. The order inside the PSNP is meaningless. If the sender keeps track of the order of LSPs sent, this indication allows a fast detection of the loss of an LSP. This MUST NOT be used to trigger faster retransmission of LSP. This MAY be used to trigger a congestion signal.

4.5. Partial SNP Interval sub-TLV

The Partial SNP Interval sub-TLV advertises the amount of time in milliseconds between periodic action for transmission of Partial Sequence Number PDUs. This time will trigger the sending of a PSNP even if the number of unacknowledged LSPs received on a given interface does not exceed LPP (Section 4.3). The time is measured from the reception of the first unacknowledged LSP.

Type: 5

Length: 2 octets

Value: partialSNPInterval in milliseconds

A node advertising this sub-TLV SHOULD send a PSNP at least once per Partial SNP Interval if one or more unacknowledged LSPs have been received on a given interface.

4.6. Operation on a LAN interface

On a LAN interface, all LSPs are link-level multicasts. Each LSP sent will be received by all ISs on the LAN and each IS will receive LSPs from all transmitters. In this section, we clarify how the flooding parameters should be interpreted in the context of a LAN.

An LSP receiver on a LAN will communicate its desired flooding parameters using a single Flooding Parameters TLV, copies of which will be received by all transmitters. The flooding parameters sent by the LSP 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 multiple transmitters that can send LSPs to the receiver LAN interface. The receiver might want to take that into account by advertising more conservative values, e.g. a higher LSP Transmission Interval. When the transmitters receive the LSP Transmission Interval value advertised by a LSP 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 LSP transmitter will receive multiple flooding parameter advertisements from different receivers that may carry different flooding parameter values. A given transmitter SHOULD use the most conservative value on a per parameter basis. For example, if the transmitter receives multiple LSP Burst Window values, it should use the smallest value.

5. Performance improvement on the receiver

This section defines two behaviors that SHOULD be implemented on the receiver.

5.1. 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.

Faster LSP flooding benefits from a faster feedback loop. This requires a reduction in the delay in sending PSNPs.

The receiver SHOULD reduce its 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, and the requirement to synchronize the LSDB more quickly. 200 ms seems to be 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 of LSPs 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, giving faster feedback to the sender. Outside of the burst case, the usual time-based PSNP approach comes into effect. The LPP SHOULD also be less than or equal to 90 as this is the maximum number of LSPs that can be

acknowledged in a PSNP at common MTU sizes, hence waiting longer would not reduce the number of PSNPs sent but would delay the acknowledgements. Based on experimental evidence, 15 unacknowledged LSPs is a good value assuming that the LSP Burst Window is at least 30 and reasonably fast CPUs for both the transmitter and receiver. More frequent PSNPs gives the transmitter more feedback on receiver progress, allowing the transmitter to continue transmitting while not burdening the receiver with undue overhead.

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.

As PSNPs also consume link bandwidth, 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.

5.2. Packet Prioritization on Receive

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

- * Hellos
- * LSPs
- * Complete Sequence Number PDUs (CSNPs) and 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.

CSNPs and PSNPs serve to trigger or acknowledge the transmission of specified LSPs. On a point-to-point link, PSNPs acknowledge the receipt of one or more LSPs. 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, the set of LSPs advertised in the PSNP cannot be marked as acknowledged and this results in needless retransmissions that will further delay transmission of other LSPs that 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 Hellos and then SNPs over LSPs. Implementations MAY also prioritize IS-IS packets over other less critical protocols.

6. Congestion and Flow Control

6.1. 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] and relies on the flow control, congestion control, and reliability mechanisms of the protocol.

Flow control creates a control loop between a transmitter and a receiver so that the transmitter does not overwhelm the receiver. TCP provides a mean for the receiver to govern the amount of data sent by the sender through the use of a sliding window.

Congestion control creates multiple interacting control loops between multiple transmitters and multiple receivers to prevent the transmitters from overwhelming the overall network. For an IS-IS adjacency, the network between two IS-IS neighbors is relatively limited in scope and consist of a link that is typically over-sized compared to the capability of the IS-IS speakers, but may also includes components inside both routers such as a switching fabric, line card CPU, and forwarding plane buffers that may experience congestion. These resources may be shared across multiple IS-IS adjacencies for the system and it is the responsibility of congestion control to ensure that these are shared reasonably.

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

The following two sections provides examples of Flow and/or Congestion control algorithms as examples that may be implemented by taking advantage of the extensions defined in this document. They are non-normative. An implementation may implement any congestion control algorithm.

6.2. Congestion and Flow Control algorithm: Example 1

6.2.1. Flow control

A flow control mechanism creates a control loop between a single instance of a transmitter and a single receiver. This example uses 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 ('rwin') indicates an allowed number of LSPs that the sender may transmit before waiting for an acknowledgment. The size of the receive window, in units of LSPs, is initialized with the value advertised by the receiver in the LSP Burst Window sub-TLV. If no value is advertised, the transmitter should initialize rwin with its own local value.

When the transmitter sends a set of LSPs to the receiver, it subtracts the number of LSPs sent from rwin. If the transmitter receives a PSNP, then rwin is incremented for each acknowledged LSP. The transmitter must ensure that the value of rwin never goes negative.

6.2.1.1. Operation on a point to point interface

By sending the LSP Burst Window sub-TLV, a node advertises to its neighbor its ability to receive that many un-acknowledged LSPs from the neighbor, with no separation interval. This is akin to a receive window or sliding window in flow control. In some implementations, this value should reflect the IS-IS socket buffer size. Special care must be taken to leave space for CSNP and PSNP (SNP) PDUs and IIHs if they share the same input queue. In this case, this document suggests advertising an LSP Burst Window corresponding to half the size of the IS-IS input queue.

By advertising an LSP Transmission Interval sub-TLV, a node advertises its ability to receive LSPs separated by at least the advertised value, outside of LSP bursts.

The LSP transmitter MUST NOT exceed these parameters. After having sent a full burst of un-acknowledged LSPs, it MUST send the following LSPs with an LSP Transmission Interval between LSP arrivals. For CPU scheduling reasons, this rate may be averaged over a small period e.g. 10 to 30ms.

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. In the worst case, an LSP is lost at the receiver and this situation is already remedied by mechanisms in [ISO10589]. After a few seconds, neighbors will exchange PSNPs (for point to point interfaces) or CSNPs (for broadcast interfaces) and recover from the lost LSPs. This worst

case should be avoided as those additional seconds impact convergence time 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.

6.2.1.2. Operation on a broadcast LAN interface

In order for the LSP Burst Window to be a useful parameter, an LSP transmitter needs to be able to keep track of the number of un-acknowledged LSPs it has sent to a given LSP receiver. On a LAN there is no explicit acknowledgment of the receipt of LSPs between a given LSP transmitter and a given LSP receiver. However, an LSP transmitter on a LAN can infer whether any LSP 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 LSP transmitter can safely set the number of un-acknowledged LSPs to zero. Since this suitable period of time is much higher than the fast acknowledgment of LSPs defined in Section 5.1, the sustainable transmission rate of LSPs will be much slower on a LAN interface than on a point to point interface. The LSP Burst Window is still very useful for the first burst of LSPs sent, especially in the case of a single node failure that requires the flooding of a relatively small number of LSPs.

6.2.2. 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.

This section describes one congestion control algorithm largely inspired by the TCP congestion control algorithm RFC 5681 [RFC5681].

The proposed algorithm uses a variable congestion window 'cwin'. It plays a role similar to the receive window described above. The main difference is that cwin is dynamically changed according to various events described below.

6.2.2.1. Core algorithm

In its simplest form, the congestion control algorithm looks like the following:

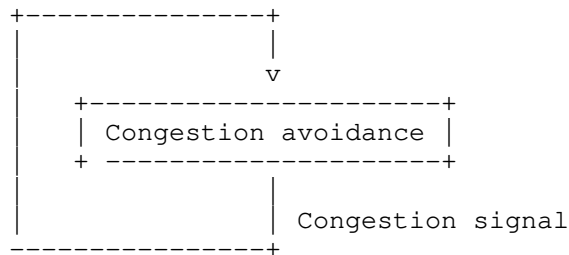


Figure 1

The algorithm starts with $cwin := LPP + 1$. In the congestion avoidance phase, $cwin$ increases as LSPs are acked: for every acked LSP, $cwin += 1 / cwin$. Thus, the sending rate roughly increases linearly with the RTT. Since the RTT is low in many IS-IS deployments, the sending rate can reach fast rates in short periods of time.

When updating $cwin$, it must not become higher than the number of LSPs waiting to be sent, otherwise the sending will not be paced by the receiving of acks. Said differently, tx pressure is needed to maintain and increase $cwin$.

When the congestion signal is triggered, $cwin$ is set back to its initial value and the congestion avoidance phase starts again.

6.2.2.2. Congestion signals

The congestion signal can take various forms. The more reactive the congestion signals, the less LSPs will be lost due to congestion. However, congestion signals too aggressive will cause a sender to keep a very low sending rate even without actual congestion on the path.

Two practical signals are given hereafter.

Timers: when receiving acknowledgements, a sender estimates the acknowledgement time of the receiver. Based on this estimation, it can infer that a packet was lost, and infer congestion on the path.

There can be a timer per LSP, but this can become costly for implementations. It is possible to use only a single timer `t1` for every LSPs: during `t1`, 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 `t1` period, 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.

There are multiple strategies to set the timeout value `t1`. It should be based on measures of the maximum acknowledgement time (MAT) of each PSNPs. The simplest one is to use a exponential moving average of the MATs, like RFC 6298 [RFC6298]. A more elaborate one is to take a running maximum of the MATs over a period of time of a few seconds. This value should include a margin of error to avoid false positives (e.g. estimated MAT measure variance) which would have a significant impact on performance.

Reordering: a sender can record its sending order and check that acknowledgements arrive on the same order than LSPs. This makes an additional assumption and should ideally be backed up by a confirmation by the receiver that this assumption stands. The O flag defined in Section 4.4 serves this purpose.

6.2.2.3. Refinement 1

With the algorithm presented above, if congestion is detected, `cwin` goes back to its initial value, and does not use the information gathered in previous congestion avoidance phases.

It is possible to use a fast recovery phase once congestion is detected, to avoid going through this linear rate of growth from scratch. When congestion is detected, a fast recovery threshold `frthresh` is set to `frthresh := cwin / 2`. In this fast recovery phase, for every acked LSP, `cwin += 1`. Once `cwin` reaches `frthresh`, the algorithm goes back to the congestion avoidance phase.

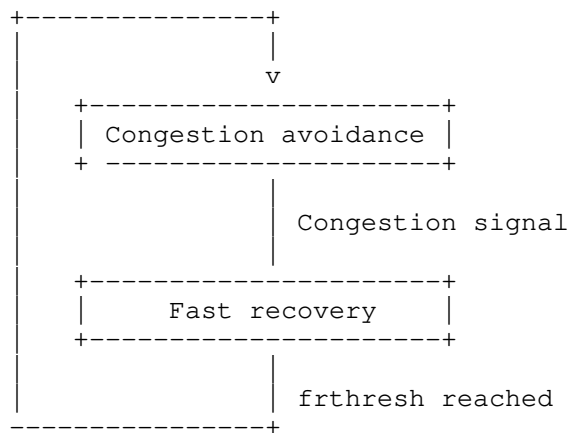


Figure 2

6.2.2.4. Refinement 2

The rates of increase were inspired from TCP RFC 5681 [RFC5681], but it is possible that a different rate of increase for $cwin$ in the congestion avoidance phase actually yields better results due to the low RTT values in most IS-IS deployments.

6.2.2.5. 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. If an LSP Burst Window is advertised, LPP SHOULD be lower and the best performance is achieved when LPP is an integer fraction of the LSP Burst Window.

Note that this congestion control algorithm benefits from the extensions proposed in this document. The advertisement of a receive window from the receiver (Section 6.2.1) avoids the use of an arbitrary maximum value by the sender. The faster acknowledgment of LSPs (Section 5.1) allows for a faster control loop and hence a faster increase of the congestion window in the absence of congestion.

6.2.3. Determining values to be advertised in the Flooding Parameters TLV

The values that a receiver advertises do not need to be perfect. If the values are too low then the transmitter will not use the full bandwidth or available CPU resources. If the values are too high then the receiver may drop some LSPs during the first RTT and this loss will reduce the usable receive window and the protocol mechanisms will allow the adjacency to recover. Flooding several orders of magnitude slower than both nodes can achieve will hurt performance, as will consistently overloading the receiver.

The values advertised need not be dynamic as feedback is provided by the acknowledgment of LSPs in SNP messages. Acknowledgments provide a feedback loop on how fast the LSPs are processed by the receiver. They also signal that the LSPs can be 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 offline 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 if the queue regularly comes close to being filled, then the 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 average 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.

6.2.4. Operation considerations

As discussed in Section 4.6, 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.

6.3. Congestion Control algorithm: Example 2

This section describes a congestion control algorithm based on performance measured by the transmitter without dependance on signaling from the receiver.

6.3.1. Router Architecture Discussion

(The following description is an abstraction - implementation details vary.)

Existing router architectures may utilize multiple input queues. On a given line card, IS-IS PDUs from multiple interfaces may be placed in a rate limited input queue. This queue may be dedicated to IS-IS PDUs or may be shared with other routing related packets.

The input queue may then pass IS-IS PDUs to a "punt queue" which is used to pass PDUs from the data plane to the control plane. The punt queue typically also has controls on its size and the rate at which packets will be punted.

An input queue in the control plane may then be used to assemble PDUs from multiple linecards, separate the IS-ISs PDU from other types of packets, and place the IS-IS PDUs in an input queue dedicated to the IS-IS protocol.

The IS-IS input queue then separates the IS-IS PDUs and directs them to an instance specific processing queue. The instance specific processing queue may then further separate the IS-IS PDUs by type (IIHs, SNPs, and LSPs) so that separate processing threads with varying priorities may be employed to process the incoming PDUs.

In such an architecture, it may be difficult for IS-IS in the control plane to accurately track the state of the various input queues and determine what value should be advertised as a current receive window.

The following section describes a congestion control algorithm based on performance measured by the transmitter without dependance on signaling from the receiver.

6.3.2. Transmitter Based Flow Control

The congestion control algorithm described in this section does not depend upon direct signaling from the receiver. Instead it adapts the transmission rate based on measurement of the actual rate of acknowledgments received.

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. It therefore incorporates the neighbor partialSNPInterval (Section 4.5) to help determine whether acknowledgments are keeping pace with the rate of LSPs transmitted. In the absence of an advertisement of partialSNPInterval a locally configured value can be used.

7. 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	y	n

Figure 3

This document creates the following sub-TLV Registry:

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

Registration Procedure(s): Expert Review

Expert(s): TBD

Reference: TBD

Type	Description
0	Reserved
1	LSP Burst Window
2	LSP Transmission Interval
3	LSPs Per PSNP
4	Flags
5	Partial SNP Interval
6-255	Unassigned

Table 1: Initial allocations

This document also requests IANA to create a new registry for assigning Flag bits advertised in the Flags sub-TLV.

Name: Flooding Parameters Flags Bits.

Registration Procedure:

Expert Review Expert(s): TBD

Bit #	Description
0	O Flag

8. Security Considerations

Security concerns for IS-IS are addressed in [ISO10589] , [RFC5304] , and [RFC5310] . These documents describe mechanisms that provide the authentication and integrity of IS-IS PDUs, including SNPs and IIHs. These authentication mechanisms are not altered by this document.

With the cryptographic mechanisms described in [RFC5304] and [RFC5310] , an attacker wanting to advertise an incorrect Flooding Parameters TLV would have to first defeat these mechanisms.

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 LSPs 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.

9. Contributors

The following people gave a substantial contribution to the content of this document and should be considered as coauthors:

Acee Lindem, Cisco Systems, acee@cisco.com

Jayesh J, Juniper Networks, jayeshj@juniper.net

10. 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.

11. References

11.1. 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, November 2002.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.

- [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", Work in Progress, Internet-Draft, draft-ietf-lsr-dynamic-flooding-09, 9 June 2021, <<https://www.ietf.org/archive/id/draft-ietf-lsr-dynamic-flooding-09.txt>>.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<https://www.rfc-editor.org/info/rfc793>>.
- [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>>.
- [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.

Appendix B. Issues for Further Discussion

[RFC Editor: Please remove this section before publication]

This section captures issues which the authors either have not yet had time to address or on which the authors have not yet reached consensus. Future revisions of this document may include new/altered text relevant to these issues.

There are no open issues at this time.

Authors' Addresses

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Les Ginsberg
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
United States of America

Email: ginsberg@cisco.com

Tony Li
Arista Networks
5453 Great America Parkway
Santa Clara, California 95054
United States of America

Email: tony.li@tony.li

Guillaume Solignac

Email: gsoligna@protonmail.com

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

Email: mkarasek@cisco.com

Chris Bowers
Juniper Networks, Inc.
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
United States of America

Email: cbowers@juniper.net

Gunter Van de Velde
Nokia
Copernicuslaan 50
2018 Antwerp
Belgium

Email: gunter.van_de_velde@nokia.com

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

Email: ppsenak@cisco.com

Tony Przygienda
Juniper
1137 Innovation Way
Sunnyvale, Ca
United States of America

Email: prz@juniper.net

Network Working Group
Internet Draft
Intended status: Standard
Expires: July 25, 2022

L. Dunbar
H. Chen
Futurewei
Aijun Wang
China Telecom
January 25, 2022

IGP Extension for 5G Edge Computing Service
draft-dunbar-lsr-5g-edge-compute-07

Abstract

This draft describes using additional site capacity and preference related metrics to influence the SPF and using Flexible Algorithms to indicate the topologies those metrics are applied. The purpose is to differentiate multiple paths with similar routing distance to one destination in 5G Local Data Network (LDN) to achieve optimal performance.

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be
accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 7, 2021.

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
(<http://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. Unbalanced Distribution due to UE Mobility.....	4
1.2. ANYCAST in 5G EC Environment.....	4
1.3. Scope of the Document.....	4
2. Conventions used in this document.....	5
3. Solution Overview.....	6
4. New Flags added to FAD Flags Sub-TLV.....	7
5. Minimum Interval for Aggregated Site Cost Advertisement	7
6. Aggregate Site Cost Advertisement in OSPF.....	8
7. "Site-Cost" Advertisement in IS-IS.....	8

8. Alternative method for Distributing Aggregated Cost.....	9
9. Manageability Considerations.....	9
10. Security Considerations.....	10
11. IANA Considerations.....	10
12. References.....	10
12.1. Normative References.....	10
12.2. Informative References.....	11
13. Appendix A:5G Edge Computing Background.....	12
13.1. Metrics to change traffic flow patterns.....	13
13.2. Reason for using IGP Based Solution.....	14
13.3. Flow Affinity to an ANYCAST server.....	15
14. Acknowledgments.....	15
1. Introduction	

In 5G Edge Computing (EC) environment, it is common for an application that needs low latency to be instantiated on multiple servers close in proximity to UEs (User Equipment). Those applications instances can be behind one or multiple application-layer load balancers. When they have relatively short flows that can go to any instance, having the cluster of them at different locations share the same IP address can minimize the impact to DNS and achieve optimal forwarding that leverages network conditions. E.g., Kubernetes for data center networking uses one single Virtual IP address for a cluster of instances of microservices so that the network can forward via multiple paths towards one single destination.

This draft describes using additional site costs to influence the shortest path computation for a specific set of prefixes. The site costs can be a group of metrics, or one aggregated cost computed based on a configured algorithm. As there are a small number of egress routers having those prefixes (or destinations) that need to incorporate site costs in SPF computation, Flexible

Algorithms [LSR-FlexAlgo] is used to indicate the need for the site costs to be considered for the specific topologies. Flexible algorithms provide mechanisms for topologies to use different IGP path algorithms.

1.1. Unbalanced Distribution due to UE Mobility

UEs' frequent moving from one 5G site to another can make it difficult to plan where the App Servers should be hosted. When a group of App servers at one location, which can be behind an application-layer load balancer, are heavily utilized, the instances for the same application at another location can be under-utilized. The difference in the routing distance to reach multiple sites where the application instances are instantiated might be relatively small in 5G LDN environment. The site capacity and preferences can be more significant than the routing distance from the application's latency and performance perspective.

Since the condition can change in days or weeks, it is difficult for the application controller to anticipate the moving and adjusting relocation of application instances.

1.2. ANYCAST in 5G EC Environment

ANYCAST is assigning the same IP address for multiple instances at different locations. Using ANYCAST can eliminate the single point of failure and bottleneck at load balancers or DNS. Another benefit is removing the dependency on how UEs resolve IP addresses for their applications. Some UEs (or clients) might use stale cached IP addresses for an extended period.

But having the same IP address at multiple locations of the 5G Edge Computing environment can be problematic because all those locations can be close in proximity. There might be a tiny difference in the routing distance to reach an application instance attached to a different edge router.

1.3. Scope of the Document

The draft is for scenarios where applications or micro services are instantiated at multiple locations behind one or multiple application layer load balancers. They have relative short flows that can go to any instances.

Under this scenario, multiple instances for the same type of services can be assigned with the same IP address, so that network condition can be utilized to achieve optimal forwarding.

From IP network perspective, application layer load balancers and app servers all appear as IP addresses. Throughout this document, the term "app server" can represent the load balancer in front of a cluster of app server instances, app server instances, or app server.

Note: for the ease of description, the EC (Edge Computing) server, Application server, App server are used interchangeably throughout this document.

2. Conventions used in this document

A-ER: Egress Edge Router to an Application Server, [A-ER] is used to describe the last router that the Application Server is attached. For 5G EC environment, the A-ER can be the gateway router to a (mini) Edge Computing Data Center.

Application Server: An application server is a physical or virtual server that hosts the software system for the application.

Application Server Location: Represent a cluster of servers at one location serving the same Application. One application may have a Layer 7 Load balancer, whose address(es) are reachable from an external IP network, in front of a set of application servers. From IP network perspective, this whole group of servers is considered as the Application server at the location.

Edge Application Server: used interchangeably with Application Server throughout this document.

EC: Edge Computing

Edge Hosting Environment: An environment providing the support required for Edge Application Server's execution.

NOTE: The above terminologies are the same as those used in 3GPP TR 23.758

Edge DC: Edge Data Center, which provides the Edge Computing Hosting Environment. It might be co-located with 5G Base Station and not only host 5G core functions, but also host frequently used Edge server instances.

gNB next generation Node B

LDN: Local Data Network

PSA: PDU Session Anchor (UPF)

SSC: Session and Service Continuity

UE: User Equipment

UPF: User Plane Function

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.

3. Solution Overview

The proposed solution is for the egress edge router (A-ER) with the application instances directly attached to

- advertise the aggregated site cost via IP prefix reachability TLV associated with the (anycast) prefix.
[note: the aggregated cost in this version of the draft is one value. Some deployment scenarios could have a set of values as the site cost.]

- use a Flag in the Flexible Algorithm TLV to indicate that the aggregated site cost needs to influence the SPF to reach the Prefix.

The aggregated site cost associated with a prefix (i.e., ANYCAST prefix) is computed based on the Capacity Index, the Preference Index, and other constraints by a consistent algorithm across all A-ERs. The capacity and preference indexes are configured to the egress routers to which the prefix is attached.

The solution assumes that the 5G EC controller or management system is aware of the ANYCAST addresses that need optimized forwarding. Only the addresses that match with the ACLs configured by the 5G EC controller will have their aggregated site cost advertised.

4. New Flags added to FAD Flags Sub-TLV

A New flag (P-flag) is added to indicate that the aggregated site cost needs to be considered for the SPF to the prefix for a specific topology. One specific topology can consist of a subset of routers within one single IGP domain.

Flags:

```
0 1 2 3 4 5 6 7...
+-+--+--+--+--+--+...
|M|P| | ...
+-+--+--+--+--+--+...
```

The detailed algorithm of integrating the routing distance and the aggregated site cost for the shortest path is out of the scope of this document.

5. Minimum Interval for Aggregated Site Cost Advertisement

The aggregated site cost associated with a prefix (e.g., an ANYCAST prefix) can be a value or a set of values configured on the router to which the prefix is attached. The aggregated site cost can be computed based on an algorithm configured on router for specific prefixes. The

detailed algorithm of computing the aggregated site cost is out of the scope of document.

As the cost change can impact the path computation, there must be a Minimum Interval for Metrics Change Advertisement which is configured on the routers to avoid route oscillations. Default is 30s.

The aggregated site cost change rate is comparable with the rate of adding or removing application instances at locations to adapt to the workload distribution changes. The rate of change could be in days or weeks. On rare occasions, there might need rate changes in hours.

6. Aggregate Site Cost Advertisement in OSPF

- IPv4: OSPFv2

A new Aggregated Cost Sub-TLV needs to be added to OSPFv2 Extended Prefix TLV [RFC7684]

- IPv6: OSPFv3

A new sub-TLV can be appended to the E-Intra-Area-Prefix-LSA, E-Inter-Area-Prefix-LSA, E-AS-External-LSA, and E-Type-7-LSA [RFC8362] to carry the Aggregated Cost.

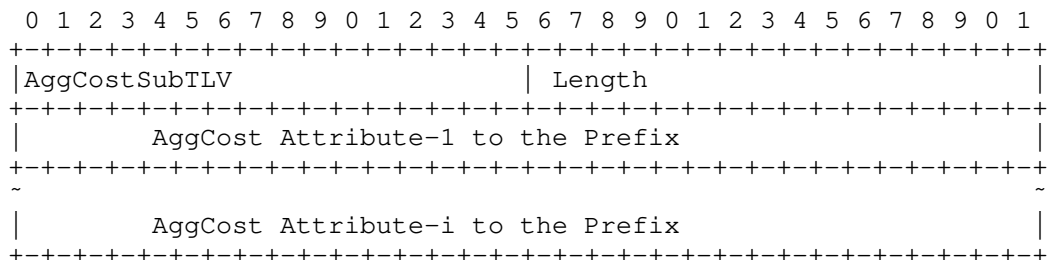


Figure 1: Aggregated cost Advertisement in OSPF

7. "Site-Cost" Advertisement in IS-IS

Aggregated Cost can be appended as subTLV to the Extended IP Reachability TLV 135 for IPv4 [RFC5305] and 236 for IPv6 [RFC5308].

For Multi-Topology with non-zero IDs, the Aggregated Cost SubTLV can be carried by Multi-topology TLV 235 for IPv4 and 237 for IPv6 [RFC5120].

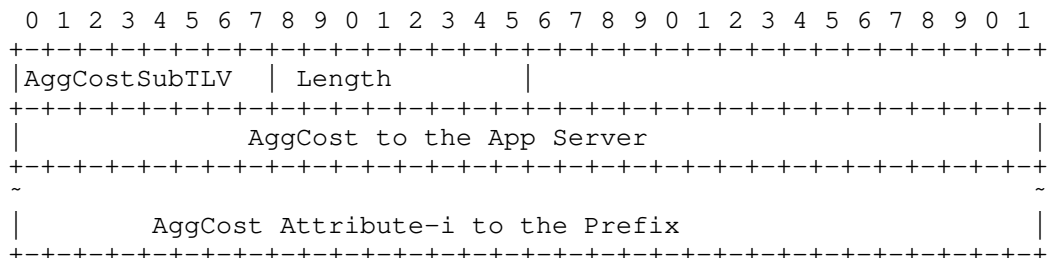


Figure 2: Aggregated cost Advertisement in IS-IS

8. Alternative method for Distributing Aggregated Cost

Section 6 and Section 7 demonstrate different ways for OSPFv2, OSPFv3, and ISIS to propagate the aggregated cost. It would be better if the aggregated cost could be advertised the same way, regardless of OSPFv2, OSPFv3, or ISIS.

Draft [draft-wang-lsr-stub-link-attributes] introduces the Stub-Link TLV for OSPFv2/v3 and ISIS protocol respectively. Considering the interfaces on an edge router that connects to the EC servers are normally configured as passive interfaces, these IP-layer App-metrics can also be advertised as the attributes of the passive/stub link. The associated prefixes can then be advertised in the "Stub-Link TLV" that is defined in [draft-wang-lsr-stub-link-attributes]. All the associated prefixes share the same characteristic of the link. Other link related sub-TLVs defined in [RFC8920] can also be attached and applied to the calculation of path to the associated prefixes."

The aggregated site cost metric can also be carried by the Stub-Link TLV defined in [draft-wang-lsr-stub-link-attributes]

9. Manageability Considerations

To be added.

10. Security Considerations

To be added.

11. IANA Considerations

The following Sub-TLV types need to be added by IANA to FlexAlgo.

- AggCostSubTLV Type for ISIS, OSPF (TBD1): IPv4 or IPv6

P-flag added to FAD Flags Sub-TLV to indicate that the Site-Cost Metrics is included in deriving Constrained IGP path to the prefix.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] J. Moy, "OSPF Version 2", RFC 2328, April 1998.
- [RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.
- [RFC7684] P. Psenak, et al, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, Nov. 2015.
- [RFC8200] S. Deering R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", July 2017.

- [RFC8326] A. Lindem, et al, "OSPFv3 Link State advertisement (LSA0 Extensibility", RFC 8362, April 2018.
- [RFC9012] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", April 2021.

12.2. Informative References

- [3GPP-EdgeComputing] 3GPP TR 23.748, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on enhancement of support for Edge Computing in 5G Core network (5GC)", Release 17 work in progress, Aug 2020.
- [5G-StickyService] L. Dunbar, J. Kaippallimalil, "IPv6 Solution for 5G Edge Computing Sticky Service", draft-dunbar-6man-5g-ec-sticky-service-00, work-in-progress, Oct 2020.
- [BGP-5G-AppMetaData] L. Dunbar, K. Majumdar, H. Wang, "BGP App Metadata for 5G Edge Computing Service", draft-dunbar-idr-5g-edge-compute-app-meta-data-03, work-in-progress, Sept 2020.
- [LSR-Flex-Algo] P. Psenak, et al, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-17, July 2021.
- [LSR-Flex-Algo-BW] S. Hegde, et al, "Flexible Algorithms: Bandwidth, Delay, Metrics and Constraints", draft-ietf-lsr-flex-algo-bw-con-01, July 2021.
- [SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.

13. Appendix A:5G Edge Computing Background

The network connecting the 5G EC servers with the 5G Base stations consists of a small number of dedicated routers that form the 5G Local Data Network (LDN) to enhance the performance of the EC services.

When a User Equipment (UE) initiates application packets using the destination address from a DNS reply or its cache, the packets from the UE are carried in a PDU session through 5G Core [5GC] to the 5G UPF-PSA (User Plan Function - PDU Session Anchor). The UPF-PSA decapsulates the 5G GTP outer header, performs NAT sometimes, before handing the packets from the UEs to the adjacent router, also known as the ingress router to the EC LDN, which is responsible for forwarding the packets to the intended destinations.

When the UE moves out of coverage of its current gNB (next-generation Node B) (gNB1), the handover procedure is initiated, which includes the 5G SMF (Session Management Function) selecting a new UPF-PSA [3GPP TS 23.501 and TS 23.502]. When the handover process is complete, the IP point of attachment is to the new UPF-PSA. The UE's IP address stays the same unless moving to different operator domain. 5GC may maintain a path from the old UPF to the new UPF for a short time for SSC [Session and Service Continuity] mode 3 to make the handover process more seamless.

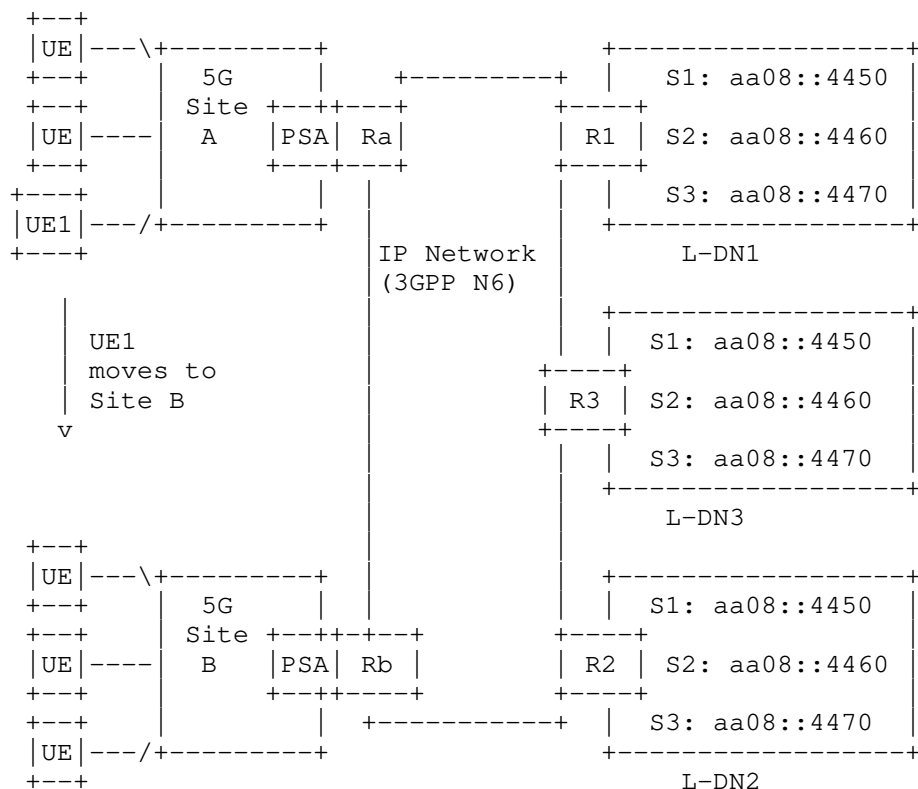


Figure 10: App Servers in different edge DCs

13.1. Metrics to change traffic flow patterns

When UEs pattern changes, the Application controller can instantiate more instances at certain locations to accommodate higher demand.

However, network layer can offer a simpler solution. By adjusting the site cost for the prefix at specific egress routers, IGP distribution of those site cost plus the flex algorithm can increase (or decrease) flows for the specific prefixes towards the certain locations.

- Capacity Index:
a numeric number, configured on all A-ERs in the domain consistently, is used to represent the capacity of an EC server attached to an A-ER. The IP addresses exposed to the A-ER can be the App Layer Load balancers that have many instances attached. At other sites, the IP address exposed is the server itself.
- Site preference index:
Is used to describe some sites are more preferred than others. For example, a site with less leasing cost has a higher preference value. Note: the preference value is configured on all A-ERs in the domain consistently by the Domain Controller.

13.2. Reason for using IGP Based Solution

IGP provides stable underlay reachability within the IGP coverage area (including hierarchy). Even though IGP has been extended to carry underlay TE or SR information, IGP has been within the core transport. IGP traditionally doesn't carry any service information.

In the networks where traditional IGP has been deployed, different addresses are for different end points. But for the 5G edge computing environment, one entity can have multiple addresses and one prefix "A" can be attached to multiple routers. E.g., one entity EAS-1 can have 3 addresses (E1/E2/E3). By adjusting the site cost for E1 on the E1 attached router (R1) for the Topology-Red, traffic destined towards E1 from the routers in the Topology-Red can be led to (or away from) to the R1. While the traffic destined towards E2/E3 stay the same.

Here are some benefits of using IGP to propagate the IP Layer App-Metrics:

- Intermediate routers can utilize the aggregated cost to reach the same prefix attached to different egress edge nodes, especially:
 - The path to the optimal egress edge node can be more accurate or shorter.

- Convergence is shorter when there is any failure along the way towards the optimal egress for the prefix.
- When there is any failure at the intended instance of the prefix, all the packets in transit can be optimally forwarded to another instance of the same prefix attached to a different egress router.
- Doesn't need the ingress nodes to establish tunnels with egress edge nodes.

There are limitations of using IGP too, such as:

- The IGP approach might not suit well to 5G EC LDN operated by multiple ISPs.
For LDN operated by multiple ISPs, BGP should be used.
[BGP-5G-AppMetaData] describes the BGP UPDATE message to propagate IP Layer App-Metrics crossing multiple ISPs.

13.3. Flow Affinity to an ANYCAST server

When multiple servers with the same IP address (ANYCAST) are attached to different A-ERs, Flow Affinity means routers sending the packets of the same flow to the same A-ER even if the cost towards the A-ER is no longer optimal.

Many commercial routers support some forms of flow affinity to ensure packets belonging to one flow be forwarded along the same path.

Editor's note: for IPv6 traffic, Flow Affinity can be achieved by routers forwarding the packets with the same Flow Label extracted from the IPv6 Header along the same path.

14. Acknowledgments

Acknowledgements to Peter Psenak, Les Ginsberg, Robert Raszuk, Acee Lindem, Shraddha Hegde, Tony Li, Gyan Mishra, Jeff Tantsura, and Donald Eastlake for their review and suggestions.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Huaimo Chen
Futurewei
Email: huaimo.chen@futurewei.com

Aijun Wang
China Telecom
Email: wangaj3@chinatelecom.cn

LSR
Internet-Draft
Intended status: Standards Track
Expires: 11 August 2022

S. Hegde
R. Bonica
C. Bowers
Juniper Networks
R. Raszuk
NTT Network Innovations
Z. Li
Huawei Technologies
D. Voyer
Bell Canada
7 February 2022

The Application Specific Link Attribute (ASLA) Any Application Bit
draft-hegde-lsr-asla-any-app-01

Abstract

RFC 8919 and RFC 8920 define Application Specific Link Attributes (ASLA). Each ASLA includes an Application Identifier Bit Mask. The Application Identifier Bit Mask includes a Standard Application Bit Mask (SABM) and a User Defined Application Bit Mask (UDABM). The SABM and UDABM determine which applications can use the ASLA as an input.

This document introduces a new bit to the Standard Application Identifier Bit Mask. This bit is called the Any Application Bit (i.e., the A-bit). If the A-bit is set, the link attribute can be used by any application. This includes currently defined applications as well as applications to be defined in the future.

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 11 August 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	2
2. Requirements Language	3
3. The Any Application Bit	3
3.1. IS-IS	4
3.2. OSPF	4
4. Backward Compatibility	4
5. Security Considerations	4
6. IANA Considerations	4
7. Acknowledgements	4
8. Normative References	4
Authors' Addresses	5

1. Introduction

[RFC8919] and [RFC8920] define Application Specific Link Attributes (ASLA). Each ASLA includes an Application Identifier Bit Mask. The Application Identifier Bit Mask includes a Standard Application Bit Mask (SABM) and a User Defined Application Bit Mask (UDABM).

Each bit in the SABM represents a standard application while each bit in the UDABM represents a user defined application. If a bit in the SABM or UDABM is set, the corresponding application can use the ASLA as an input. If a bit in the SABM or UDABM is not set, the corresponding application cannot use the associated ASLA as an input.

According to [RFC8919]:

"If link attributes are advertised associated with zero-length Application Identifier Bit Masks for both standard applications and user-defined applications, then any standard application and/or any user-defined application is permitted to use that set of link attributes so long as there is not another set of attributes

advertised on that same link that is associated with a non-zero-length Application Identifier Bit Mask with a matching Application Identifier Bit set."

This restriction introduces complexity. For example, assume that a network runs many applications. All applications use Attribute 1 as an input. So, it would be convenient to advertise Attribute 1 with a zero-length SABM / UDABM.

However, Applications X and Y also use Attribute 2 as an input. Because Applications X and Y required unique values for Attribute 2, Attribute 2 cannot be advertised with a zero-length SABM. Therefore, Attribute 1 cannot be advertised with a zero-length SABM / UDABM either, because Applications X and Y require it. This would result in having to set the application X and application Y bits on attribute 1 in the entire network on each link and is operationally complex.

This document reduces operational complexity by introducing a new bit to the Standard Application Identifier Bit Mask. This bit is called the Any Application Bit (i.e., the A-bit). If the A-bit is set, the link attribute can be used by any application. This includes currently defined applications as well as applications to be defined in the future.

2. 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.

3. The Any Application Bit

A new bit is defined in the Standard Application Identifier Bit Mask. This bit is called the Any Application Bit (i.e., the A-bit). If the A-bit is set, the link attribute can be used by any application. This includes currently defined applications as well as applications to be defined in the future.

If a link advertises an ASLA twice, once with the A-bit set and once with a more specific Application Identifier Bit set, the indicated application MUST use the value from the ASLA with the more specific Application Indicator Bit set.

3.1. IS-IS

IS-IS uses Bit 4 of the SABM to encode the A-bit.

3.2. OSPF

OSPF uses Bit 4 of the SABM to encode the A-bit.

4. Backward Compatibility

The solution described in this document is backward compatible with [RFC8919] and [RFC8920]. An implementation that does not recognize the A-bit will process the SABM as specified in [RFC8919] and [RFC8920].

Implementations MAY advertise attributes under both A bit and with SABM and UDABM length set to zero for backward compatibility reasons. When same attributes are received with A bit set as well as in ASLA with SABM and UDABM set to zero, the attributes MUST be used from the ASLA with SABM and UDABM set to zero and procedures described in RFC 8919 sec 6.2 MUST be followed.

5. Security Considerations

The security considerations discussed in [RFC8919] and [RFC8920] are applicable to this document. This document does not introduce any new security risks.

6. IANA Considerations

This document requests that IANA add the following entry to the registry titled "Link Attribute Application Identifiers" under the "Interior Gateway Protocol (IGP) Parameters" registry:

- * Bit: 4
- * Name: Any Application (A-bit)
- * Reference: This document

7. Acknowledgements

TBD

8. 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>>.
- [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>>.
- [RFC8919] Ginsberg, L., Psenak, P., Previdi, S., Henderickx, W., and J. Drake, "IS-IS Application-Specific Link Attributes", RFC 8919, DOI 10.17487/RFC8919, October 2020, <<https://www.rfc-editor.org/info/rfc8919>>.
- [RFC8920] Psenak, P., Ed., Ginsberg, L., Henderickx, W., Tantsura, J., and J. Drake, "OSPF Application-Specific Link Attributes", RFC 8920, DOI 10.17487/RFC8920, October 2020, <<https://www.rfc-editor.org/info/rfc8920>>.

Authors' Addresses

Shraddha Hegde
Juniper Networks
Exora Business Park
Bangalore 560103
KA
India

Email: shraddha@juniper.net

Ron Bonica
Juniper Networks
2251 Corporate Park Drive
Herndon, Virginia 20171
United States of America

Email: rbonica@juniper.net

Chris Bowers
Juniper Networks

Email: cbowers@juniper.net

Robert Raszuk
NTT Network Innovations

Email: robert@raszuk.net

Zenbin
Huawei Technologies

Email: lizhenbin@huawei.com

Dan Voyer
Bell Canada

Email: daniel.voyer@bell.ca

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 22 April 2022

Z. Hu
S. Peng
X. Xi
Huawei
19 October 2021

IGP Extensions for Path MTU
draft-hu-lsr-igp-path-mtu-00

Abstract

Segment routing (SR) leverages the source routing mechanism. It allows for a flexible definition of end-to-end paths within IGP topologies by encoding paths as sequences of topological sub-paths which are called segments. These segments are advertised by the link-state routing protocols (IS-IS and OSPF). Unlike the MPLS, SR does not have the specific path construction signaling so that it cannot support the Path MTU. This draft provides the necessary IS-IS and OSPF extensions about the Path MTU that need to be used on SR. Here, the term "OSPF" means both OSPFv2 and OSPFv3.

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 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 22 April 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. Terminology	3
3. IGP Extension	4
3.1. IS-IS Extension	4
3.2. OSPF Extension	5
4. Acknowledgements	6
5. IANA Considerations	6
6. Security Considerations	7
7. References	7
Authors' Addresses	8

1. Introduction

Segment routing (SR) leverages the source routing mechanism. SR allows for a flexible definition of end-to-end paths within IGP topologies by encoding paths as sequences of topological sub-paths which are called segments. These segments are advertised by the link-state routing protocols (IS-IS and OSPF). The SR architecture as well as the routing policy is proposed in [RFC8402] and [I-D.ietf-spring-segment-routing-policy]. Two types of segments are defined, Prefix segments and Adjacency segments. Prefix segments represent an ECMP-aware shortest path to a prefix (or a node), as per the state of the IGP topology. Adjacency segments represent a hop over a specific adjacency between two nodes in the IGP. A prefix segment is typically a multi-hop path while an adjacency segment, in most of the cases, is a one-hop path. SR can compute the paths from end to end and without requiring any LDP or RSVP-TE signaling. SR supports per-flow explicit routing while just maintaining per-flow state only at the source node.

SR architecture supports the distributed scenario and the centralized scenario. In the distributed scenario, the segments are allocated and signaled by IGP or BGP and a node needs to compute the source-routed policy. Some necessary IS-IS and OSPF extensions for SR are proposed in [RFC8665] [RFC8666] [RFC8667]. In a centralized scenario, the SR controller decides which nodes need to steer which packets on which source-routed policies. However, in both conditions, the MTU is not included in the SR policy. As the SR may push more MPLS labels or SRv6 SIDs in the packet header, the packets are more likely to be larger than the minimum MTU in the path compared to the traditional MPLS forwarding process. Unfortunately, with the current mechanisms in SR, the path MTU information cannot be obtained in advance. Therefore it cannot be ensured that the packet size is less than the path MTU which is the minimum link MTU of all the links in a path between a source node and a destination node. The definition of the path MTU is discussed in [RFC1191] [RFC8201].

This draft describes the necessary IS-IS and OSPF extensions for obtaining the path MTU to be used on SR. New sub-TLVs are introduced for both the IS-IS and OSPF protocols. With the IGP flooding process in the distributed scenario or the BGP transmission to the controller, the ingress node or the controller is able to compute the path MTU for the SR policy.

2. Terminology

Router: A node that forwards IP packets not explicitly addressed to itself.

Interface: A node's attachment to a link.

Segment: An instruction a node executes on the incoming packet. For example, forward packet according to shortest path to destination or a specific interface, etc..

SR Policy: An ordered list of segments.

MTU: Maximum Transmission Unit, the size in bytes of the largest IP packet, including the IP header and payload, that can be transmitted on a link or path. Note that this could more properly be called the IP MTU, to be consistent with how other standards organizations use the acronym MTU.

Link MTU: The maximum transmission unit, i.e., maximum IP packet size in bytes, that can be conveyed in one piece over a link. Be aware that this definition is different from the definition used by other standards organizations.

For IETF documents, link MTU is uniformly defined as the IP MTU over the link. This includes the IP header, but excludes link layer headers and other framing that is not part of IP or the IP payload.

Be aware that other standards organizations generally define link MTU to include the link layer headers.

For the MPLS data plane, this size includes the IP header and data (or other payload) and the label stack but does not include any lower-layer headers. A link may be an interface (such as Ethernet or Packet-over-SONET), a tunnel (such as GRE or IPsec), or an LSP.

Path: The set of links traversed by a packet between a source node and a destination node

Path MTU: The minimum link MTU of all the links in a path between a source node and a destination node.

3. IGP Extension

This document describes IS-IS and OSPF extensions to flood the router interface MTU to each node within an IGP domain. Then the controller or the original node collects all the link MTUs from the routers. So the original node can compute the minimum link MTU of all the links in the path. The source node can limit the packet size less than the path MTU.

3.1. IS-IS Extension

A new sub-TLV called link MTU sub-TLV is defined for TLVs 22, 23, 25, 141, 222, 223 in the Router Information LSP to carry the MTU of the interface associated with the link. Each sub-TLV is encoded as shown in Figure 1.

Type: MTU, 1 byte, TBD.

Length: # of octets in the value field, 1 byte.

Value: The value is the MTU size of a link, 2 bytes.

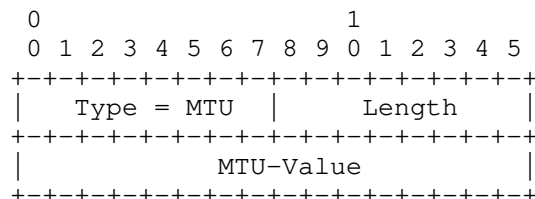


Figure 1: Figure 1: Link MTU Sub-TLV for the IS-IS extension

The use and meaning of these fields are as follows:

Type - A single octet encoding the sub-TLV type. Here the type is 1 octet.

Length - A single octet encoding the total length of the value field of the sub-TLV in octets.

MTU-Value - Two octets encoding the MTU size of the sub-TLV. This field identifies the size of the router interfaces.

This sub-TLV is optional.

This document defines a link MTU sub-TLV for IS-IS extension. The codepoints need to be determined by the IANA.

3.2. OSPF Extension

A new sub-TLV called link MTU sub-TLV is defined in the corresponding LSA as specified for OSPFv2 and OSPFv3 to carry the MTU of the interface associated with the link. Each sub-TLV is encoded as shown in Figure 2.

Type: MTU, 2 bytes, TBD.

Length: # of octets in the value field, 2 bytes.

Value: The value is the MTU size of a link, 2 bytes.

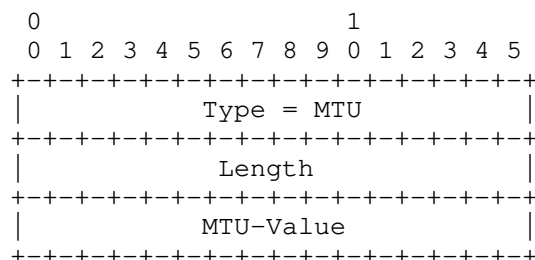


Figure 2: Figure 2: Link MTU Sub-TLV for the OSPF extension

The use and meaning of these fields are as follows:

Type - Two octets encoding the TLV type. Here the type is 2 octets.

For OSPFv2, the link MTU is advertised as an optional sub-TLV of the OSPFv2 Extended Link TLV in the OSPFv2 Extended Link Opaque LSA as defined in [RFC7684] and the codepoints need to be determined by the IANA.

For OSPFv3, the link MTU is advertised as an optional sub-TLV of the Router-Link TLV in the OSPFv3 E-Router-LSA as defined in [RFC8362] and the codepoints need to be determined by the IANA.

Length - Two octets encoding the total length of the value field of the sub-TLV in octets.

MTU-Value - Two octets encoding the MTU size of the TLV. This field identifies the size of the router interfaces.

If the link MTU sub-TLV is advertised for multiple times for the same link in different OSPFv2 Extended Link Opaque LSAs or OSPFv3 E-Router-LSAs originated by the same OSPF router, the link MTU sub-TLV in the OSPFv2 Extended Link Opaque LSA with the smallest Opaque ID or in the OSPFv3 E-Router-LSA with the smallest Link State ID MUST be used by receiving OSPF routers.

4. Acknowledgements

TBD.

5. IANA Considerations

This document requests that IANA allocates a new sub-TLV type as defined in Section 3.1 from the "Sub-TLVs for TLVs 22, 23, 25, 141, 222, and 223 (Extended IS reachability, IS Neighbor Attribute, L2 Bundle Member Attributes, inter-AS reachability information, MT-ISN, and MT IS Neighbor Attribute TLVs)" registry as specified.

Value	Description	Reference
TBD	IS-IS Link MTU	This document

Figure 3: Figure 3: IS-IS Link MTU

This document requests that IANA allocates a new sub-TLV type as defined in Section 3.2 from the "OSPFv2 Extended Link TLV Sub-TLVs" registry.

Value	Description	Reference
TBD	OSPFv2 Link MTU	This document

Figure 4: Figure 4: OSPFv2 Link MTU

This document requests that IANA allocates a new sub-TLV type as defined in Section 3.2 from the "OSPFv3 Extended LSA Sub-TLVs" registry.

Value	Description	Reference
TBD	OPSFv3 Link MTU	This document

Figure 5: Figure 5: OSPFv3 Link MTU

6. Security Considerations

These extensions to IS-IS and OSPF do not add any new security issues to the existing IGP.

7. References

- [I-D.ietf-spring-segment-routing-policy]
 Filts, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", Work in Progress, Internet-Draft, draft-ietf-spring-segment-routing-policy-13, 28 May 2021, <<https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-13.txt>>.
- [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/info/rfc1191>>.
- [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>>.
- [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>>.
- [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>>.

- [RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/info/rfc8201>>.
- [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>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.
- [RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <<https://www.rfc-editor.org/info/rfc8665>>.
- [RFC8666] Psenak, P., Ed. and S. Previdi, Ed., "OSPFv3 Extensions for Segment Routing", RFC 8666, DOI 10.17487/RFC8666, December 2019, <<https://www.rfc-editor.org/info/rfc8666>>.
- [RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <<https://www.rfc-editor.org/info/rfc8667>>.

Authors' Addresses

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

Email: huzhibo@huawei.com

Shuping Peng
Huawei
Huawei Bld., No. 156 Beiqing Rd.
Beijing
100095
China

Email: pengshuping@huawei.com

Xing Xi
Huawei
Huawei Bld., No. 156 Beiqing Rd.
Beijing
100095
China

Email: xixing1@huawei.com

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.

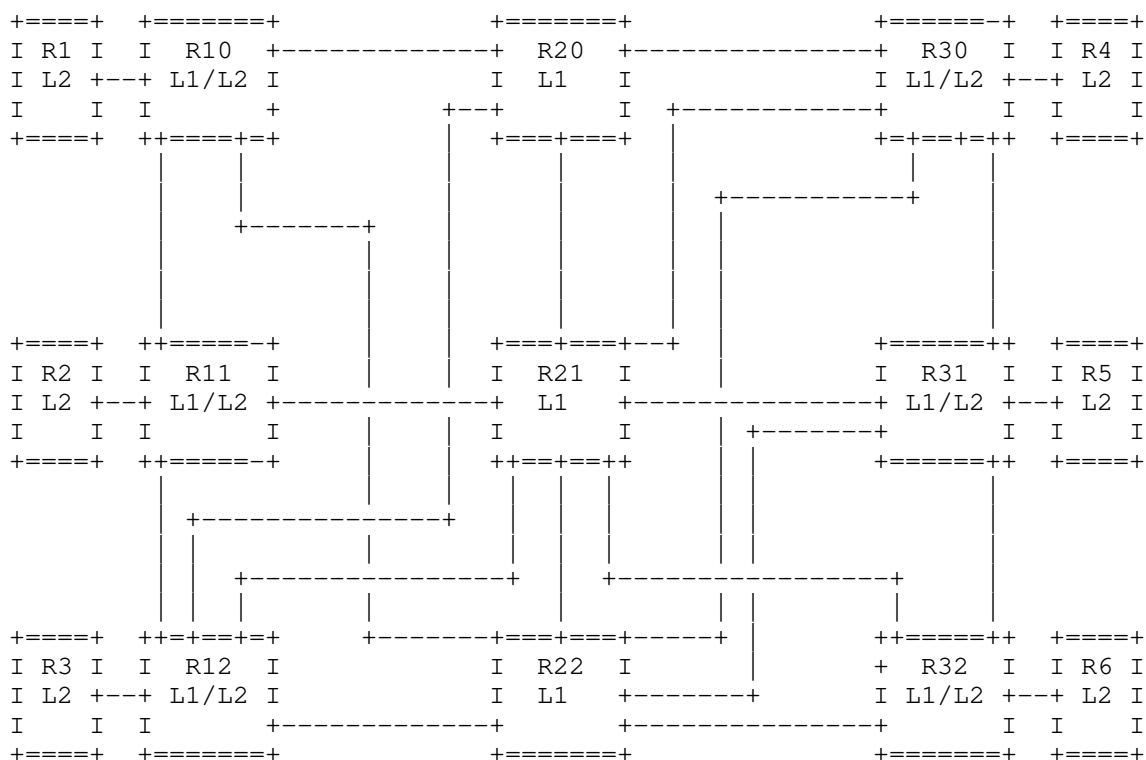


Figure 1: Example Topology of L1 with L2 Borders

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.

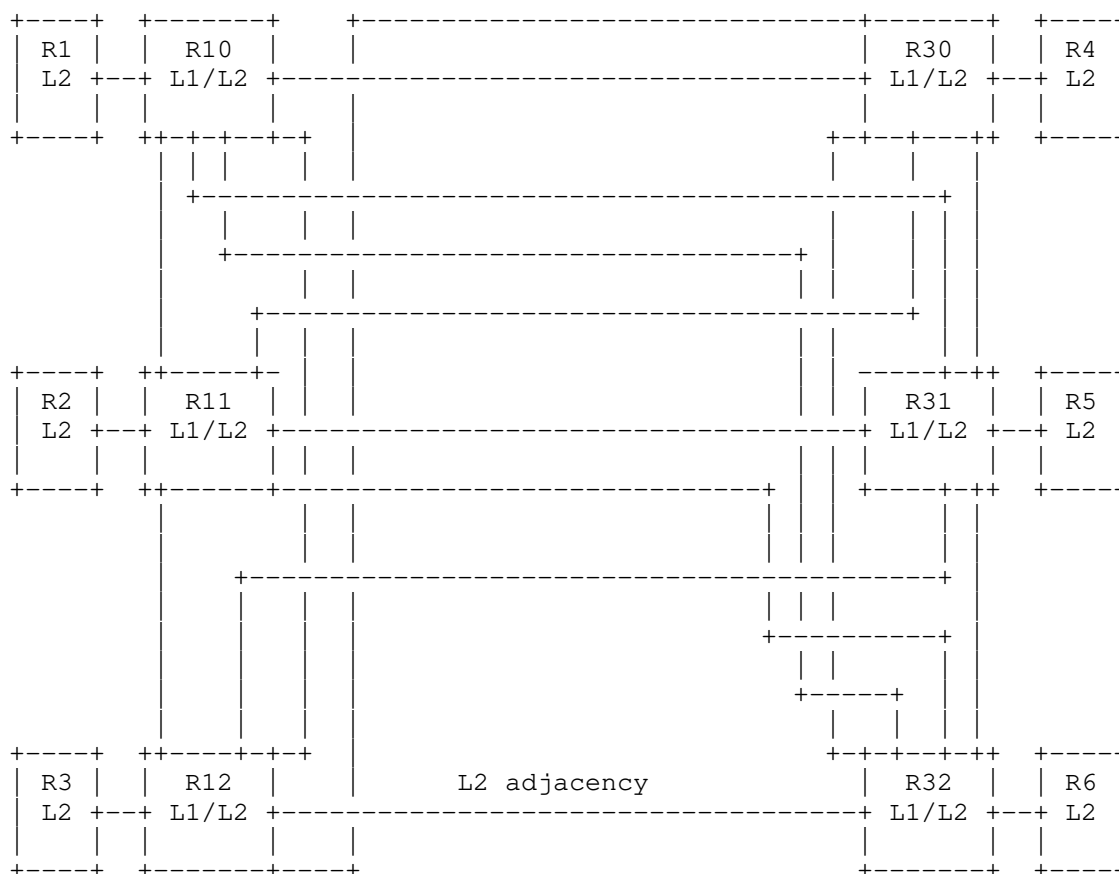


Figure 2: Example topology represented in L2 with a full mesh of L2 adjacencies between L1/L2 nodes

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 $R \cdot n$, where R is the number of flood reflectors used and n is the number of L1/L2 nodes. This compares quite favorably with $n \cdot (n-1)/2$ L2 adjacencies required in a topologically fully meshed L2 solution.

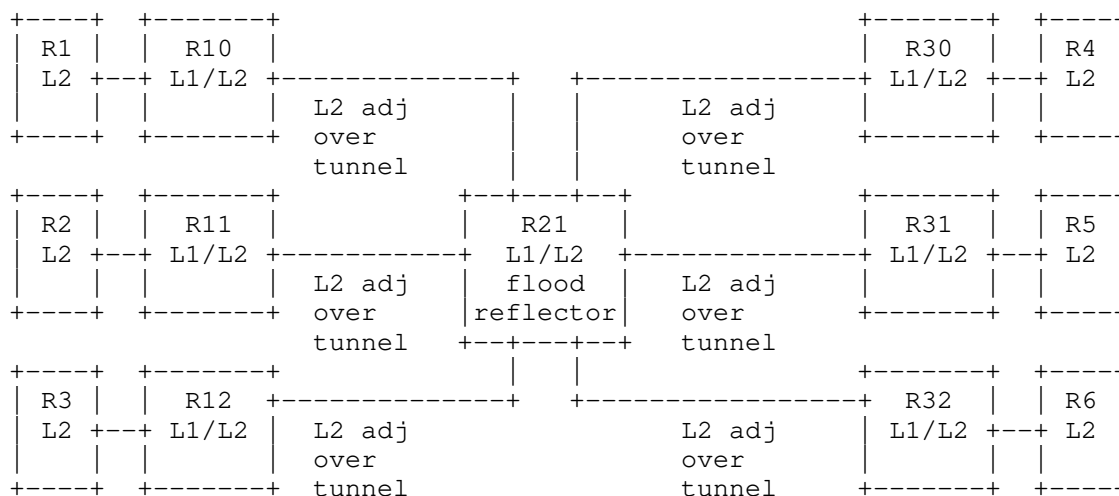


Figure 3: Example topology represented in L2 with L2 adjacencies from each L1/ L2 node to a single flood reflector

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.

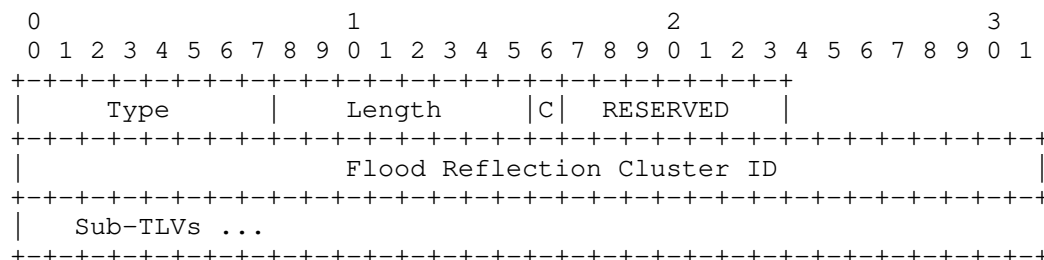
Another possible variation is for an implementation to approximate with the tunnel cost the cost of the underlying topology.

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.

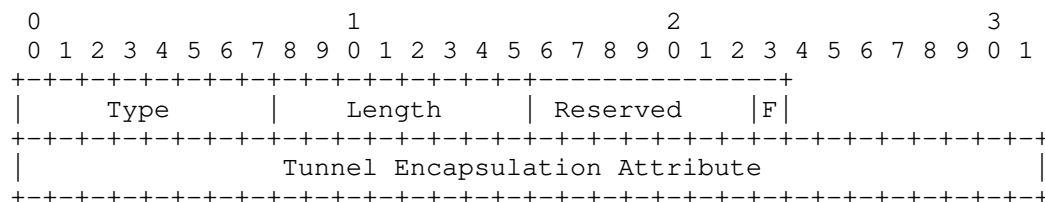
Flood Reflection Cluster ID: The Flood Reflection Cluster Identifier

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

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

Email: cbowers@juniper.net

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

Email: 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

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 22, 2022

C. Li
G. Xu
Z. Hu
Z. Zhou
Huawei
October 19, 2021

IS-IS Extensions for Link Bit Error Ratio
draft-li-lsr-isis-link-ber-00

Abstract

In certain networks, network-performance criteria (e.g., latency) are becoming as critical to data-path selection as other metrics. This document describes extensions to IS-IS Traffic Engineering (TE) Metric Extensions (RFC 8570). This draft provides the necessary IS-IS extensions about the link bit error ratio (LBER) that need to be used to describe network-performance.

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 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 April 22, 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. LBER Extensions to IS-IS	3
3. Sub-TLV Details	3
3.1. Unidirectional Link BIT ERROR RATIO Sub-TLV	3
4. Announcement Thresholds and Filters	4
5. Announcement Suppression	4
6. Network Stability and Announcement Periodicity	4
7. Enabling and Disabling Sub-TLVs	4
8. Compatibility	4
9. Acknowledgements	4
10. IANA Considerations	4
11. Security Considerations	5
12. References	5
Authors' Addresses	5

1. Introduction

In certain networks, network-performance criteria (e.g., latency) are becoming as critical to data-path selection as other metrics. This document describes extensions to IS-IS Traffic Engineering (TE) Metric Extensions (RFC 8570). This draft provides the necessary IS-IS extensions about the link bit error ratio (LBER) that need to be used to describe network-performance. A new sub-TLV is introduced for IS-IS.

As other IS-IS TE Metric Extensions (e.g., unidirectional link loss, unidirectional link delay), Unidirectional link bit error ratio described in this document is also meant to be used as part of the operation of the routing protocol to enhance Constrained Shortest Path First (CSPF), or for other uses such as supplementing the data used by the controller to compute the policy path.

2. LBER Extensions to IS-IS

This document registers a new IS-IS TE sub-TLV in the "Sub-TLVs for TLVs 22, 23, 141, 222, and 223" registry. This new sub-TLV provides ways to distribute LBER.

This document registers a sub-TLV:

Type	Description
TBD	Unidirectional Link BIT ERROR RATIO

Figure 1

The new sub-TLV include a bit called the Anomalous (or "A") bit. When the A bit is clear (or when the sub-TLV does not include an A bit), the sub-TLV describes steady-state link performance.

3. Sub-TLV Details

3.1. Unidirectional Link BIT ERROR RATIO Sub-TLV

This sub-TLV advertises the bit error ratio between two directly connected IS-IS neighbors. The link bit error ratio advertised by this sub-TLV MUST be the packet bit error from the local neighbor to the remote neighbor (i.e., the forward-path loss). The format of this sub-TLV is shown in the following diagram:

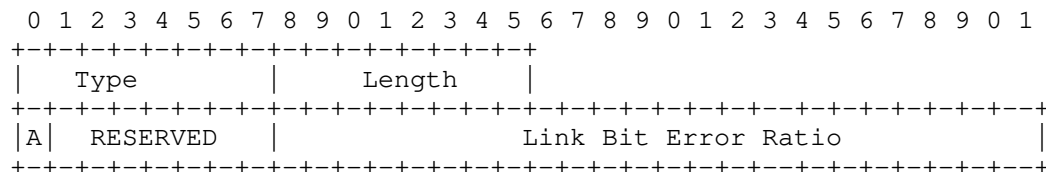


Figure 2: Unidirectional Link BIT ERROR RATIO Sub-TLV for the IS-IS extension

Type: TBD (suggested value 40) is to be assigned by IANA.

Length: 4.

A bit: This field represents the Anomalous (A) bit. The A bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A bit is cleared, the sub-TLV represents steady-state link performance.

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

Link Bit Error Ratio: This 24-bit field carries Link Bit Error Ratio as a percentage of the total traffic sent over a configurable interval. The basic unit is 0.000003%, where $(2^{24} - 2)$ is 50.331642%. This value is the highest link bit error percentage that can be expressed (the assumptions being that (1) precision is more important on high-speed links than the ability to advertise link bit error ratio greater than this and (2) high-speed links with over 50% bit error are unusable). Therefore, measured values that are larger than the field maximum SHOULD be encoded as the maximum value.

This sub-TLV is optional.

4. Announcement Thresholds and Filters

This document uses the same principle for announcement thresholds and filters as described in RFC 8570.

5. Announcement Suppression

This document uses the same principle for announcement suppression as described in RFC 8570.

6. Network Stability and Announcement Periodicity

This document uses the same principle for network stability and announcement periodicity as described in RFC 8570.

7. Enabling and Disabling Sub-TLVs

Implementations MUST make it possible to enable or disable each sub-TLV based on configuration.

8. Compatibility

Unrecognized sub-TLVs should be silently ignored.

9. Acknowledgements

TBD.

10. IANA Considerations

This document requests that IANA allocates new sub-TLV types as defined in Section 2 from the "Sub-TLVs for TLVs 22, 23, 25, 141, 222, and 223 (Extended IS reachability, IS Neighbor Attribute, L2

Bundle Member Attributes, inter-AS reachability information, MT-ISN, and MT IS Neighbor Attribute TLVs)" registry as specified.

Value	Description	Reference
TBD	Unidirectional LBER	This document

Figure 3: Unidirectional LBER

11. Security Considerations

These extensions to IS-IS do not add any new security issues to the existing IGP.

12. References

- [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>>.
- [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

Chenxi Li
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lichenxi1@huawei.com

Guoqi Xu
Huawei
Huawei Bld., No. 156 Beiqing Rd.
Beijing 100095
China

Email: xuguoqi@huawei.com

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

Email: huzhibo@huawei.com

Tianran Zhou
Huawei
Huawei Bld., No. 156 Beiqing Rd.
Beijing 100095
China

Email: zhoutianran@huawei.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:

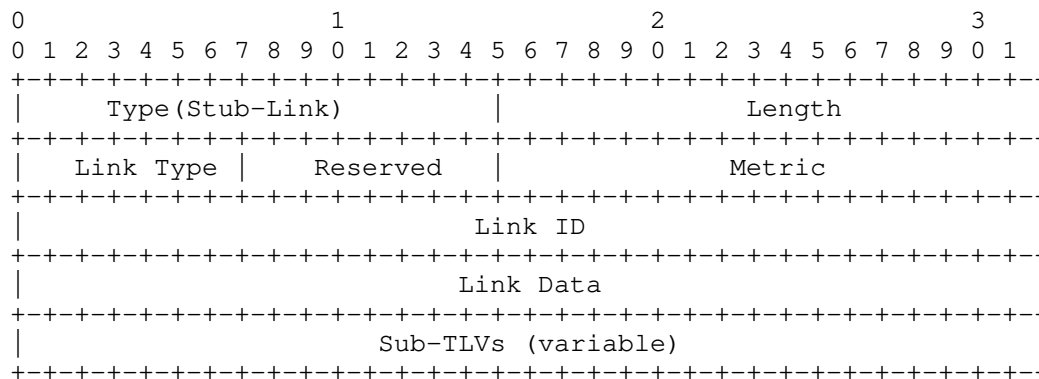


Figure 1: OSPFv2 Extended Stub-Link TLV

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.

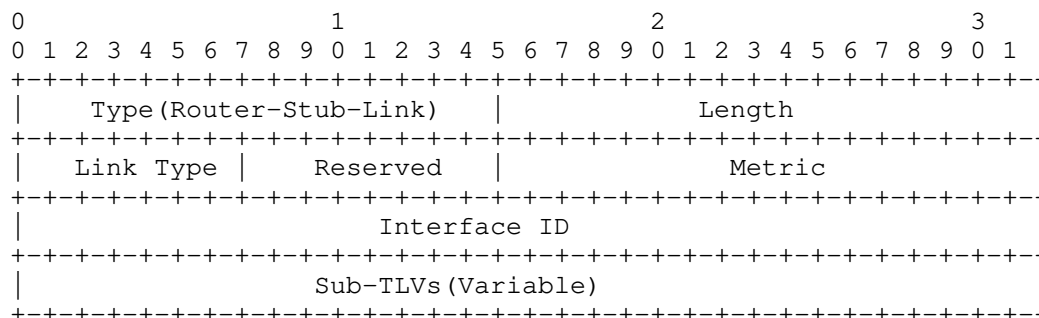


Figure 2: OSPFv3 Router-Stub-Link TLV

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:

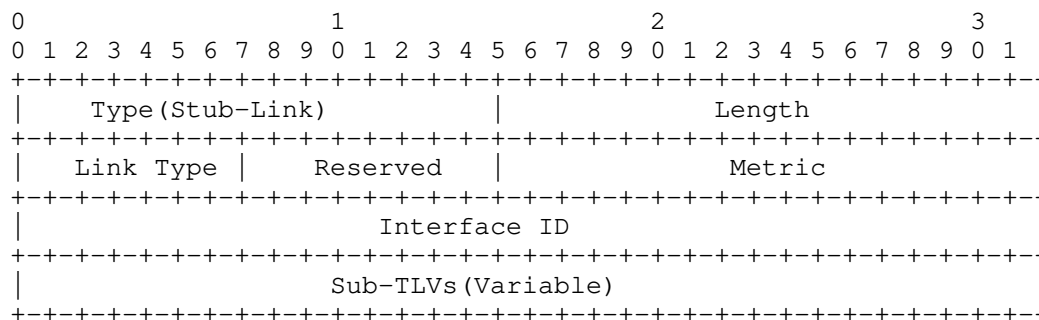


Figure 3: ISIS Stub-Link TLV

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:

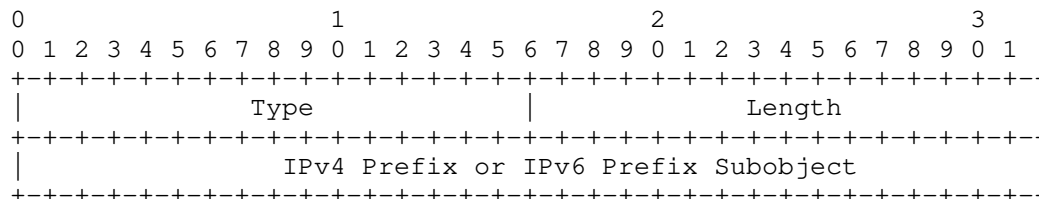


Figure 4: Stub-Link Prefix Sub-TLV

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.

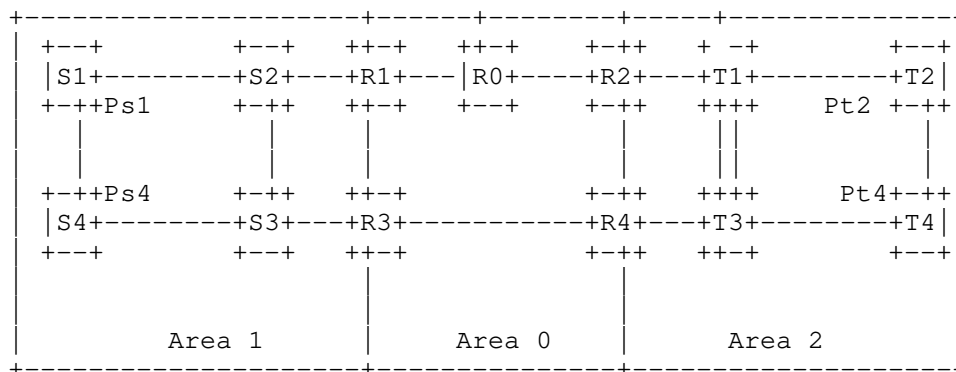


Figure 1: OSPF Inter-Area Prefix Unreachable Announcement Scenario

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.

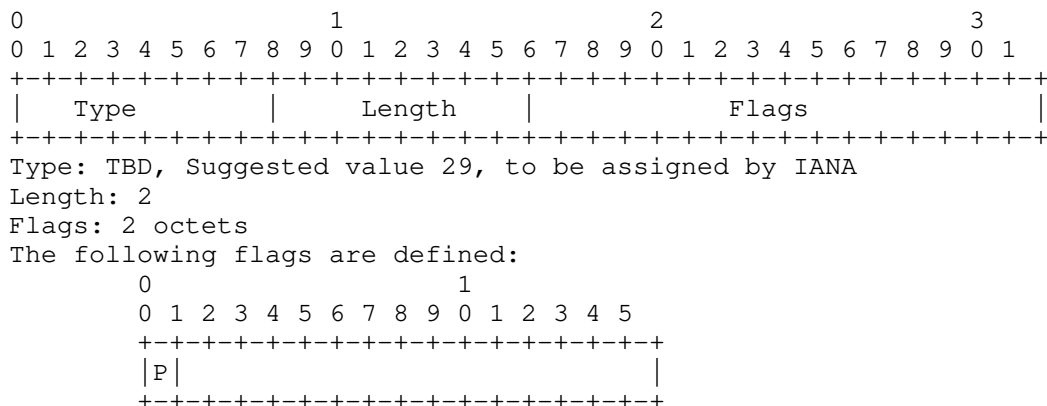
PUAM is now able to provide the negative prefix component flooded across the backbone to the other areas along with the summary prefix, which is now immediately programmed into the RIB control plane. MPLS LSP exact match or SRv6 LPM match over fail over path can now be established to the alternate egress 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

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

Email: xiaoyaqun@huawei.com

LSR Working Group
Internet-Draft
Intended status: Standards Track
Expires: 19 February 2022

X. Min
Z. Zhang
ZTE Corp.
W. Cheng
China Mobile
18 August 2021

Signaling Flow-ID Label Capability and Flow-ID Readable Label Depth
Using IGP and BGP-LS
draft-xzc-lsr-mpls-flc-flrd-01

Abstract

Flow-ID Label (FL) is used for MPLS flow identification and flow-based performance measurement with alternate marking method. The ability to process Flow-ID labels is called Flow-ID Label Capability (FLC), and the capability of reading the maximum label stack depth and performing FL-based performance measurement is called Flow-ID Readable Label Depth (FRLD). This document defines a mechanism to signal the FLC and the FRLD using IGP and BGP-LS.

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 19 February 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
1.1. Terminology	2
2. Advertising FLC Using IGP	3
2.1. Advertising FLC Using IS-IS	3
2.2. Advertising FLC Using OSPFv2	3
2.3. Advertising FLC Using OSPFv3	4
3. Advertising FRLD Using IGP	4
4. Signaling FLC and FRLD in BGP-LS	4
5. Security Considerations	5
6. IANA Considerations	5
7. Acknowledgements	5
8. Normative References	5
Authors' Addresses	7

1. Introduction

As specified in [I-D.ietf-mpls-inband-pm-encapsulation], Flow-ID Label (FL) is used for MPLS flow identification and flow-based performance measurement with alternate marking method.

Flow-ID Label may appear multiple times in a label stack with variable depth, so both the Flow-ID Label Capability (FLC) and the Flow-ID Readable Label Depth (FRLD) are defined in [I-D.ietf-mpls-inband-pm-encapsulation].

Analogous to [RFC9088] and [RFC9089], this document defines a mechanism to signal the FLC and the FRLD using IGP and BGP-LS, specifically, IGP includes IS-IS, OSPFv2 and OSPFv3.

1.1. Terminology

This memo makes use of the terms defined in [I-D.ietf-mpls-inband-pm-encapsulation] and [RFC8491].

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.

2. Advertising FLC Using IGP

Even though FLC is a property of the node, in some cases it is advantageous to associate and advertise the FLC with a prefix, so FLC is advertised with a prefix in this document.

If a router has multiple interfaces, the router **MUST NOT** announce FLC unless all of its interfaces are capable of processing FLs.

If the router supports FLs on all of its interfaces, it **SHOULD** advertise the FLC with every local host prefix it advertises in IGP.

2.1. Advertising FLC Using IS-IS

Next to the ELC Flag (E-flag) defined in Section 3 of [RFC9088], a new bit FLC Flag (F-flag) is defined, which is Bit 4 in the Prefix Attribute Flags [RFC7794], as shown in Figure 1.

```

      0 1 2 3 4 5 6 7...
    +--+--+--+--+--+--+...
    |X|R|N|E|F|      ...
    +--+--+--+--+--+--+...

```

Figure 1: Prefix Attribute Flags

F-Flag: FLC Flag (Bit 4)

Set for local host prefix of the originating node if it supports FLC on all interfaces.

The FLC signaling **MUST** be preserved when a router propagates a prefix between ISIS levels [RFC5302].

2.2. Advertising FLC Using OSPFv2

Next to the ELC Flag (E-flag) defined in Section 3.1 of [RFC9089], a new bit FLC Flag (F-flag) is defined, which is Bit 3 in Flags field of OSPFv2 Extended Prefix TLV [RFC7684]:

0x10 - F-Flag (FLC Flag): Set for local host prefix of the originating node if it supports FLC on all interfaces.

The FLC signaling **MUST** be preserved when an OSPFv2 Area Border Router (ABR) distributes information between areas. To do so, an ABR **MUST** originate an OSPFv2 Extended Prefix Opaque LSA [RFC7684] including the received FLC setting.

2.3. Advertising FLC Using OSPFv3

Next to the ELC Flag (E-flag) defined in Section 3.2 of [RFC9089], a new bit FLC Flag (F-flag) is defined, which is Bit 0 in OSPFv3 PrefixOptions field [RFC5340]:

0x80 - F-Flag (FLC Flag): Set for local host prefix of the originating node if it supports FLC on all interfaces.

The FLC signaling MUST be preserved when an OSPFv3 Area Border Router (ABR) distributes information between areas. The setting of the FLC Flag in the Inter-Area-Prefix-LSA [RFC5340] or in the Inter-Area-Prefix TLV [RFC8362], generated by an ABR, MUST be the same as the value the FLC Flag associated with the prefix in the source area.

3. Advertising FRLD Using IGP

As requested by [RFC8491], IANA has created an IANA-managed registry titled "IGP MSD-Types" to identify MSD-Types. A new MSD-Type, called FRLD-MSD, is defined to advertise the FRLD of a given router. The MSD-Type code 3 is requested to be assigned by IANA for FRLD-MSD. The MSD-Value field is set to the FRLD in the range between 0 to 255.

If a router has multiple interfaces with different capabilities of reading the maximum label stack depth, the router MUST advertise the smallest value found across all of its interfaces.

For IS-IS, the FRLD is advertised in a Node MSD Sub-TLV [RFC8491] using the FRLD-MSD type.

For OSPF including both OSPFv2 and OSPFv3, the FRLD is advertised in a Node MSD TLV [RFC8476] using the FRLD-MSD type.

The absence of FRLD-MSD advertisements indicates only that the advertising node does not support advertisement of this capability.

4. Signaling FLC and FRLD in BGP-LS

The IGP extensions defined in this document can be advertised via BGP-LS (Distribution of Link-State and TE Information Using BGP) [RFC7752] using existing BGP-LS TLVs.

The FLC is advertised using the Prefix Attribute Flags TLV as defined in [RFC9085].

The FRLD-MSD is advertised using the Node MSD TLV as defined in [RFC8814].

5. Security Considerations

This document does not raise any additional security issues beyond those of the specifications referred to in the list of normative references.

6. IANA Considerations

This document requests the following allocations from IANA:

- Bit 4 in the Bit Values for Prefix Attribute Flags Sub-TLV registry is requested to be assigned to the FLC Flag (F-Flag).
- Flag 0x10 in the OSPFv2 Extended Prefix TLV Flags registry is requested to be assigned to the FLC Flag (F-Flag).
- Bit 0x80 in the "OSPFv3 Prefix Options (8 bits)" registry is requested to be assigned to the FLC Flag (F-Flag).
- Type 3 in the IGP MSD-Types registry is requested to be assigned to the FLRD-MSD.

7. Acknowledgements

TBA.

8. Normative References

- [I-D.ietf-mpls-inband-pm-encapsulation]
Cheng, W., Min, X., Zhou, T., Dong, X., and Y. Peleg,
"Encapsulation For MPLS Performance Measurement with
Alternate Marking Method", Work in Progress, Internet-
Draft, draft-ietf-mpls-inband-pm-encapsulation-01, 11
April 2021, <<https://www.ietf.org/archive/id/draft-ietf-mpls-inband-pm-encapsulation-01.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>>.
- [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>>.
- [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>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [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>>.
- [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>>.
- [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>>.
- [RFC8476] Tantsura, J., Chunduri, U., Aldrin, S., and P. Psenak, "Signaling Maximum SID Depth (MSD) Using OSPF", RFC 8476, DOI 10.17487/RFC8476, December 2018, <<https://www.rfc-editor.org/info/rfc8476>>.
- [RFC8491] Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg, "Signaling Maximum SID Depth (MSD) Using IS-IS", RFC 8491, DOI 10.17487/RFC8491, November 2018, <<https://www.rfc-editor.org/info/rfc8491>>.
- [RFC8814] Tantsura, J., Chunduri, U., Talaulikar, K., Mirsky, G., and N. Triantafyllis, "Signaling Maximum SID Depth (MSD) Using the Border Gateway Protocol - Link State", RFC 8814, DOI 10.17487/RFC8814, August 2020, <<https://www.rfc-editor.org/info/rfc8814>>.

- [RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <<https://www.rfc-editor.org/info/rfc9085>>.
- [RFC9088] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using IS-IS", RFC 9088, DOI 10.17487/RFC9088, August 2021, <<https://www.rfc-editor.org/info/rfc9088>>.
- [RFC9089] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using OSPF", RFC 9089, DOI 10.17487/RFC9089, August 2021, <<https://www.rfc-editor.org/info/rfc9089>>.

Authors' Addresses

Xiao Min
ZTE Corp.
Nanjing
China

Email: xiao.min2@zte.com.cn

Zheng(Sandy) Zhang
ZTE Corp.
Nanjing
China

Email: zhang.zheng@zte.com.cn

Weiqiang Cheng
China Mobile
Beijing
China

Email: chengweiqiang@chinamobile.com