ISIS Working Group
Internet Draft
Intended status: Standards Track
Expires: December 2016

Dave Allan, Uma Chunduri
Ericsson

June 2016

IS-IS extensions for Computed Multicast applied to MPLS based Segment Routing
draft-allan-isis-spring-multicast-00

Abstract

This document describes the IS-IS extensions required to support multicast for MPLS based Segment Routing. In this approach IS-IS speakers compute their role in multicast tree construction based on the information in the IS-IS routing information base.

Status of this Memo

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

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/1id-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 December 2016.

Table of Contents

1. Introduction

[ALLAN-1] describes a solution for multicast for Segment Routing with MPLS data plane in which source specific multicast distribution trees (MDTs) are computed from information distributed via an IGP. Using this approach, both any-source multicast (ASM) and engineered p2mp trees can be supported.

This memo describes TLVs for IS-IS to support the segment routing multicast approach as described in [ALLAN-1].

1.1. Authors

David Allan, Uma Chunduri

1.2. 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 RFC2119 [RFC2119].

2. Conventions used in this document

2.1. Terminology

Candidate replication point - is a node will potentially need to install state as determined at some intermediate step in MDT computation. It will either resolve to having no role or a role as a replication point once multicast has converged.

Candidate role - refers to any potential combination of roles on a given MDT as determined at some intermediate step in MDT computation. For example, a node with a candidate role may be a leaf and may be a candidate replication point.

Downstream - refers to the direction along the shortest path to one or more leaves for a given multicast distribution tree

Multicast convergence - is when all computation and state installation to ensure the FIB reflects the multicast information in the IGP is complete.

Pinned path - Is a unique shortest path extending from a leaf upstream towards the root for a given MDT. Therefore is a component of an MDT that must be there. It will not necessarily extend from the leaf all the way to the root during intermediate computation steps. A pinned path can result from pruning operations.

Role - refers specifically to a node that is either a root, a leaf or a replication node for a given multicast distribution tree

Unicast convergence- is when all computation and state installation to ensure the FIB reflects the unicast information in the IGP is complete.

Upstream - refers to the direction along the shortest path to the root of a given multicast distribution tree

3. Overview

[ALLAN-1] adds the concept of the multicast segment to the Segment Routing architecture [IDSR].

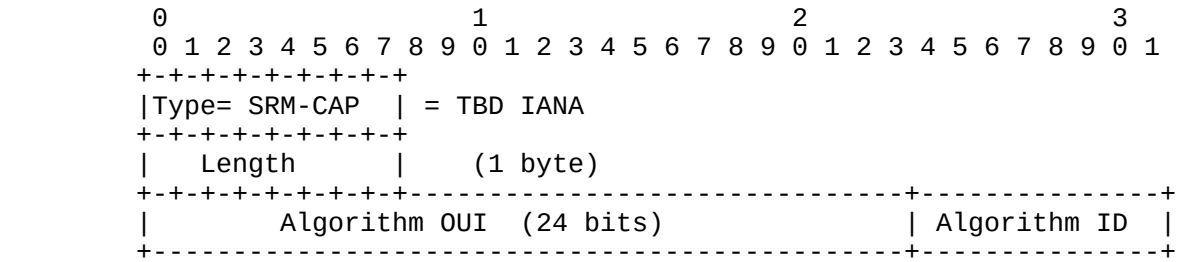
Extending the IS-IS to support multicast segments adds synchronization of knowledge of: multicast SIDs, multicast group membership and agreement on the algorithm to use for computation of multicast distribution trees(MDTs) across the set of IS-IS speakers in an area/domain. This document specifies the TLVs necessary for IS-IS to support multicast segments in the Segment Routing architecture.

4. New TLVs

4.1. Compute Capability TLV

The presence of this sub-TLV in an LSP (TLV 144 defined in [RFC6329]) indicates both that the originating node supports computed spring multicast, and the algorithm that is configured to be used for a particular topology. All nodes supporting computed multicast are required to agree on the algorithm for correct operation of the network for that topology.

The format of the sub-TLV is:



Where:

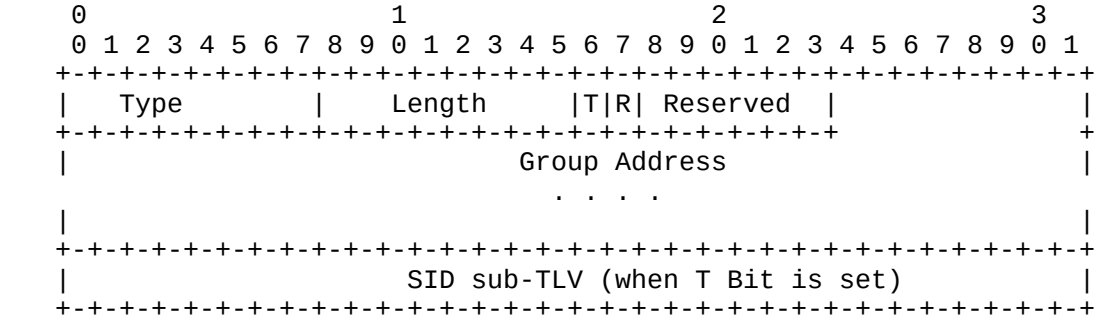
The upper 24 bits contains an organizationally unique identifier (as per [RFC7042]) and the lower 8 bits contains an algorithm identifier.

The default algorithm supported as per [ALLAN-1] is identified by Algorithm OUI = 0x008037 (Ericsson), Algorithm ID = 0x01 (default).

4.2. SRM SID Multicast Group Binding sub-TLV

The SID Multicast Group Binding sub-TLV communicates the binding between the SID specific to the MDT for the multicast group originating at the advertising node and the multicast group address as well as transmit and receive interest for the advertising node. Note that if the TLV does not have the T bit set, the SID TLV is not included in the message. The encoding is as a sub-TLV from the 135,

235, 236 and 237 registry. This sub-TLV SHOULD be advertised when N bit set in the IPv4/IPv6 Extended Reachability Attribute Flags for the corresponding prefix as defined in [RFC7794]. The encoding of the SID sub-TLV is as per section 2.3 of [SPRING-ISIS].



Where:

Type = TBD (IANA assignment from TLV 135, 235, 236 and 237 registry)

Length = 8 octet Variable which includes size of the T/R, reserved fields, multicast group address and the SID TLV

T-bit indicates that this node is a source for the multicast group specified in the sub-tlv.

R-bit indicates that this node is a receiver for the multicast group specified in the sub-tlv.

Group address = 4 octet IPv4 multicast group address (when used with TLV 135 or 235), 16 octet IPv6 multicast group address (when used with TLV 236 or 237).

SID sub-TLV contains the segment ID to use for this multicast segment.

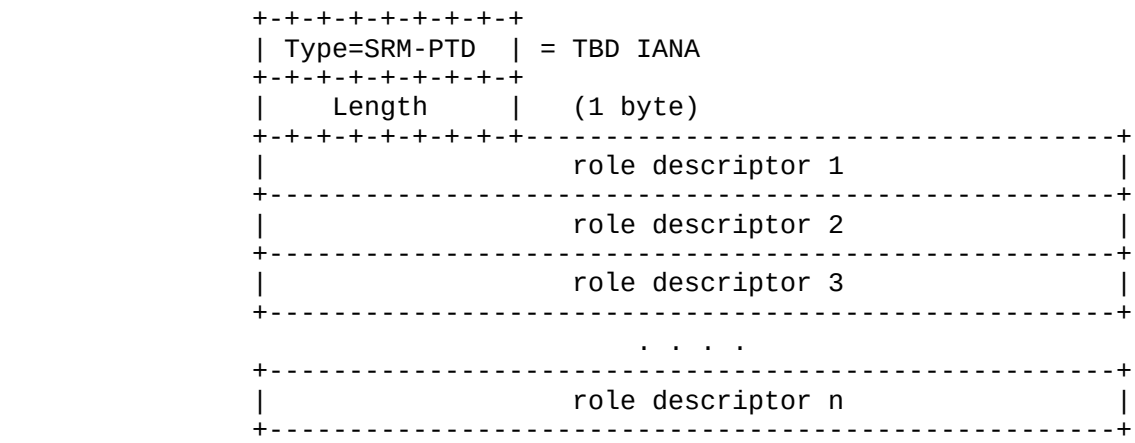
4.3. SRM Pinned Tree Descriptor sub-TLV

The pinned tree descriptor defines all nodes that have a role in a multicast distribution tree, and their relationship to the individual multicast segments that define the tree. The encoding is an unstructured list, where if the tree description exceeds 252 bytes, it may simply use more than one sub-TLV. This sub-TLV SHOULD be

advertised when N bit set in the IPv4/IPv6 Extended Reachability Attribute Flags for the corresponding prefix as defined in [RFC7794].

The encoding of a role descriptor is in the form of upstream_SID/unicast_SID/downstream _SID, where each of the SIDs is encoded as a sub-TLV as per [SPRING-ISIS].

The root of the MDT (and originator of the TLV) will have a NULL upstream SID, transit waypoints will have both a defined upstream and downstream multicast segment SID, and a leaf will have a NULL downstream SID. The unicast SID corresponds to the node for which the entry defines its role.

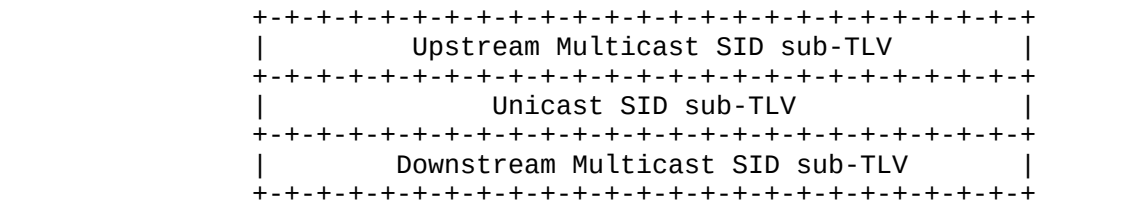


Where:

Type = TBD (IANA assignment from TLV 135, 235, 236 and 237 registry)

Length = Variable and represents size of the role descriptors

The encoding of a role descriptor is in the form:



5. Acknowledgements

6. Security Considerations

For a future version of this document.

7. IANA Considerations

This memo requires the allocation of:

1) a value for each of the SRM SID Multicast Group Binding sub-TLV, and the SRM Pinned Tree Descriptor sub-TLV from the "Sub-TLVs for TLVs 135, 235, 236, and 237" registry.

2) A value for the SRM Capability sub-TLV from the "Sub-TLVs for TLV 144" registry.

8. References

8.1. Normative References

[IS-IS] ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC6329] Fedyk et.al. "IS-IS Extensions Supporting IEEE 802.1aq Shortest Path Bridging", IETF RFC 6329, April 2012

[RFC7042] Eastlake, D. et. al.,"IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", IETF RFC 7042, October 2013

[RFC7794] Ginsberg et. al., "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", IETF RFC 7794, March 2016

[SPRING-ISIS] Previdi el.al. "IS-IS Extensions for Segment Routing", IETF work in progress, draft-ietf-isis-segment-routing-extensions-06, December 2015

8.2. Informative References

[IDSR] Filsfils et.al., "Segment Routing Architecture", IETF work in progress, draft-ietf-spring-segment-routing-08, May 2016

[ALLAN-1] Allan et.al., "A Framework for Computed Multicast applied to MPLS based Segment Routing", draft-allan-spring-mpls-mcast-framework-01, July 2016

9. Authors' Addresses

Dave Allan (editor)
Ericsson
300 Holger Way
San Jose, CA 95134
USA
Email: david.i.allan@ericsson.com

Uma Chunduri
Ericsson
300 Holger Way
San Jose, CA 95134
USA
Email: uma.chunduri@ericsson.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: July 8, 2017

M. Chen
Huawei
L. Ginsberg
S. Previdi
Cisco Systems
D. Xiaodong
China Mobile

January 4, 2017

ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering
draft-chen-isis-rfc5316bis-02

Abstract

This document describes extensions to the ISIS (ISIS) protocol to support Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering (TE) for multiple Autonomous Systems (ASes). It defines ISIS-TE extensions for the flooding of TE information about inter-AS links, which can be used to perform inter-AS TE path computation.

No support for flooding information from within one AS to another AS is proposed or defined in this document.

This document obsoletes [RFC5316]

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 http://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 July 8, 2017.

Table of Contents

1. Introduction

[RFC5305] defines extensions to the ISIS protocol [RFC1195] to support intra-area Traffic Engineering (TE). The extensions provide a way of encoding the TE information for TE-enabled links within the network (TE links) and flooding this information within an area. The extended IS reachability TLV and traffic engineering router ID TLV, which are defined in [RFC5305], are used to carry such TE information. The extended IS reachability TLV has several nested sub-TLVs that describe the TE attributes for a TE link.

[RFC6119] and [RFC5307] define similar extensions to ISIS in support of IPv6 and Generalized Multiprotocol Label Switching (GMPLS) TE respectively.

Requirements for establishing Multiprotocol Label Switching (MPLS) TE Label Switched Paths (LSPs) that cross multiple Autonomous Systems (ASes) are described in [RFC4216]. As described in [RFC4216], a method SHOULD provide the ability to compute a path spanning multiple ASes. So a path computation entity that may be the head-end Label Switching Router (LSR), an AS Border Router (ASBR), or a Path Computation Element (PCE) [RFC4655] needs to know the TE information not only of the links within an AS, but also of the links that connect to other ASes.

In this document, a new TLV, which is referred to as the inter-AS reachability TLV, is defined to advertise inter-AS TE information, three new sub-TLVs are defined for inclusion in the inter-AS reachability TLV to carry the information about the remote AS number and remote ASBR ID. The sub-TLVs defined in [RFC5305][RFC6119] and other documents for inclusion in the extended IS reachability TLV for describing the TE properties of a TE link are applicable to be included in the Inter-AS Reachability TLV for describing the TE properties of an inter-AS TE link as well. Also, two more new sub-TLVs are defined for inclusion in the IS-IS router capability TLV to carry the TE Router ID when the TE Router ID needs to reach all routers within an entire ISIS routing domain. The extensions are equally applicable to IPv4 and IPv6 as identical extensions to [RFC5305] and [RFC6119]. Detailed definitions and procedures are discussed in the following sections.

This document does not propose or define any mechanisms to advertise any other extra-AS TE information within ISIS. See Section 2.1 for a full list of non-objectives for this work.

2. Problem Statement

As described in [RFC4216], in the case of establishing an inter-AS TE LSP that traverses multiple ASes, the Path message [RFC3209] may include the following elements in the Explicit Route Object (ERO) in order to describe the path of the LSP:

- a set of AS numbers as loose hops; and/or
- a set of LSRs including ASBRs as loose hops.

Two methods for determining inter-AS paths are currently being discussed. The per-domain method [RFC5152] determines the path one domain at a time. The backward recursive method [RFC5441] uses cooperation between PCEs to determine an optimum inter-domain path. The sections that follow examine how inter-AS TE link information could be useful in both cases.

2.1. A Note on Non-Objectives

It is important to note that this document does not make any change to the confidentiality and scaling assumptions surrounding the use of ASes in the Internet. In particular, this document is conformant to the requirements set out in [RFC4216].

The following features are explicitly excluded:

- There is no attempt to distribute TE information from within one AS to another AS.
- There is no mechanism proposed to distribute any form of TE reachability information for destinations outside the AS.
- There is no proposed change to the PCE architecture or usage.
- TE aggregation is not supported or recommended.
- There is no exchange of private information between ASes.
- No ISIS adjacencies are formed on the inter-AS link.

2.2. Per-Domain Path Determination

In the per-domain method of determining an inter-AS path for an MPLS-TE LSP, when an LSR that is an entry-point to an AS receives a Path message from an upstream AS with an ERO containing a next hop that is an AS number, it needs to find which LSRs (ASBRs) within the local AS are connected to the downstream AS. That way, it can compute a TE

LSP segment across the local AS to one of those LSRs and forward the Path message to that LSR and hence into the next AS. See Figure 1 for an example.

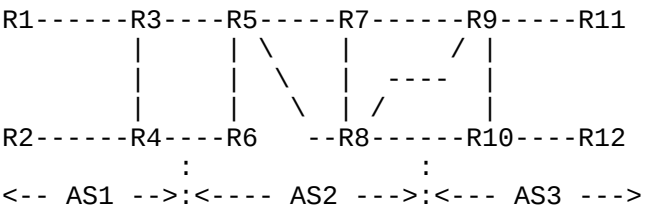


Figure 1: Inter-AS Reference Model

The figure shows three ASes (AS1, AS2, and AS3) and twelve LSRs (R1 through R12). R3 and R4 are ASBRs in AS1. R5, R6, R7, and R8 are ASBRs in AS2. R9 and R10 are ASBRs in AS3.

If an inter-AS TE LSP is planned to be established from R1 to R12, the AS sequence will be: AS1, AS2, AS3.

Suppose that the Path message enters AS2 from R3. The next hop in the ERO shows AS3, and R5 must determine a path segment across AS2 to reach AS3. It has a choice of three exit points from AS2 (R6, R7, and R8), and it needs to know which of these provide TE connectivity to AS3, and whether the TE connectivity (for example, available bandwidth) is adequate for the requested LSP.

Alternatively, if the next hop in the ERO is the entry ASBR for AS3 (say R9), R5 needs to know which of its exit ASBRs has a TE link that connects to R9. Since there may be multiple ASBRs that are connected to R9 (both R7 and R8 in this example), R5 also needs to know the TE properties of the inter-AS TE links so that it can select the correct exit ASBR.

Once the Path message reaches the exit ASBR, any choice of inter-AS TE link can be made by the ASBR if not already made by the entry ASBR that computed the segment.

More details can be found in Section 4 of [RFC5152], which clearly points out why advertising of inter-AS links is desired.

To enable R5 to make the correct choice of exit ASBR, the following information is needed:

- List of all inter-AS TE links for the local AS.
- TE properties of each inter-AS TE link.

- o AS number of the neighboring AS connected to by each inter-AS TE link.

- o Identity (TE Router ID) of the neighboring ASBR connected to by each inter-AS TE link.

In GMPLS networks, further information may also be required to select the correct TE links as defined in [RFC5307].

The example above shows how this information is needed at the entry-point ASBRs for each AS (or the PCEs that provide computation services for the ASBRs). However, this information is also needed throughout the local AS if path computation functionality is fully distributed among LSRs in the local AS, for example to support LSPs that have start points (ingress nodes) within the AS.

2.3. Backward Recursive Path Computation

Another scenario using PCE techniques has the same problem. [RFC5441] defines a PCE-based TE LSP computation method (called Backward Recursive Path Computation) to compute optimal inter-domain constrained MPLS-TE or GMPLS LSPs. In this path computation method, a specific set of traversed domains (ASes) are assumed to be selected before computation starts. Each downstream PCE in domain(i) returns to its upstream neighbor PCE in domain(i-1) a multipoint-to-point tree of potential paths. Each tree consists of the set of paths from all boundary nodes located in domain(i) to the destination where each path satisfies the set of required constraints for the TE LSP (bandwidth, affinities, etc.).

So a PCE needs to select boundary nodes (that is, ASBRs) that provide connectivity from the upstream AS. In order for the tree of paths provided by one PCE to its neighbor to be correlated, the identities of the ASBRs for each path need to be referenced. Thus, the PCE must know the identities of the ASBRs in the remote AS that are reached by any inter-AS TE link, and, in order to provide only suitable paths in the tree, the PCE must know the TE properties of the inter-AS TE links. See the following figure as an example.

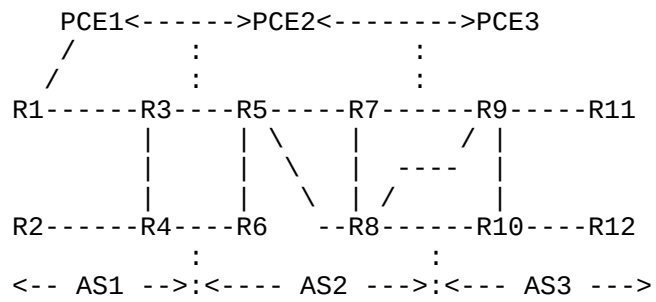


Figure 2: BRPC for Inter-AS Reference Model

The figure shows three ASes (AS1, AS2, and AS3), three PCEs (PCE1, PCE2, and PCE3), and twelve LSRs (R1 through R12). R3 and R4 are ASBRs in AS1. R5, R6, R7, and R8 are ASBRs in AS2. R9 and R10 are ASBRs in AS3. PCE1, PCE2, and PCE3 cooperate to perform inter-AS path computation and are responsible for path segment computation within their own domain(s).

If an inter-AS TE LSP is planned to be established from R1 to R12, the traversed domains are assumed to be selected: AS1->AS2->AS3, and the PCE chain is: PCE1->PCE2->PCE3. First, the path computation request originated from the PCC (R1) is relayed by PCE1 and PCE2 along the PCE chain to PCE3. Then, PCE3 begins to compute the path segments from the entry boundary nodes that provide connection from AS2 to the destination (R12). But, to provide suitable path segments, PCE3 must determine which entry boundary nodes provide connectivity to its upstream neighbor AS (identified by its AS number), and must know the TE properties of the inter-AS TE links. In the same way, PCE2 also needs to determine the entry boundary nodes according to its upstream neighbor AS and the inter-AS TE link capabilities.

Thus, to support Backward Recursive Path Computation, the same information listed in Section 2.2 is required. The AS number of the neighboring AS connected to by each inter-AS TE link is particularly important.

3. Extensions to ISIS-TE

Note that this document does not define mechanisms for distribution of TE information from one AS to another, does not distribute any form of TE reachability information for destinations outside the AS, does not change the PCE architecture or usage, does not suggest or recommend any form of TE aggregation, and does not feed private information between ASes. See Section 2.1.

In this document, for the advertisement of inter-AS TE links, a new TLV, which is referred to as the inter-AS reachability TLV, is defined. Three new sub-TLVs are also defined for inclusion in the inter-AS reachability TLV to carry the information about the neighboring AS number and the remote ASBR ID of an inter-AS link. The sub-TLVs defined in [RFC5305], [RFC6119], and other documents for inclusion in the extended IS reachability TLV are applicable to be included in the inter-AS reachability TLV for inter-AS TE links advertisement. Also, two other new sub-TLVs are defined for inclusion in the IS-IS router capability TLV to carry the TE Router ID when the TE Router ID is needed to reach all routers within an entire ISIS routing domain.

While some of the TE information of an inter-AS TE link may be available within the AS from other protocols, in order to avoid any dependency on where such protocols are processed, this mechanism carries all the information needed for the required TE operations.

3.1. Inter-AS Reachability TLV

The inter-AS reachability TLV has type 141 (see Section 6.1) and contains a data structure consisting of:

```
4 octets of Router ID
3 octets of default metric
1 octet of control information, consisting of:
   1 bit of flooding-scope information (S bit)
   1 bit of up/down information (D bit)
   6 bits reserved
1 octet of length of sub-TLVs
0-246 octets of sub-TLVs, where each sub-TLV consists of a sequence of:
      1 octet of sub-type
      1 octet of length of the value field of the sub-TLV
      0-244 octets of value
```

Compared to the extended reachability TLV which is defined in [RFC5305], the inter-AS reachability TLV replaces the "7 octets of System ID and Pseudonode Number" field with a "4 octets of Router ID" field and introduces an extra "control information" field, which consists of a flooding-scope bit (S bit), an up/down bit (D bit), and 6 reserved bits.

The Router ID field of the inter-AS reachability TLV is 4 octets in length, which contains the IPv4 Router ID of the router who generates the inter-AS reachability TLV. The Router ID SHOULD be identical to the value advertised in the Traffic Engineering Router ID TLV [RFC5305]. If no Traffic Engineering Router ID is assigned, the Router ID SHOULD be identical to an IP Interface Address [RFC1195]

advertised by the originating IS. If the originating node does not support IPv4, then the reserved value 0.0.0.0 MUST be used in the Router ID field and the IPv6 Router ID sub-TLV MUST be present in the inter-AS reachability TLV. The Router ID could be used to indicate the source of the inter-AS reachability TLV.

The flooding procedures for inter-AS reachability TLV are identical to the flooding procedures for the GENINFO TLV, which are defined in Section 4 of [RFC6823]. These procedures have been previously discussed in [RFC4971]. The flooding-scope bit (S bit) SHOULD be set to 0 if the flooding scope is to be limited to within the single IGP area to which the ASBR belongs. It MAY be set to 1 if the information is intended to reach all routers (including area border routers, ASBRs, and PCEs) in the entire ISIS routing domain. The choice between the use of 0 or 1 is an AS-wide policy choice, and configuration control SHOULD be provided in ASBR implementations that support the advertisement of inter-AS TE links.

The sub-TLVs defined in [RFC5305], [RFC6119], and other documents for describing the TE properties of a TE link are also applicable to the inter-AS reachability TLV for describing the TE properties of an Inter-AS TE link. Apart from these sub-TLVs, four new sub-TLVs are defined for inclusion in the inter-AS reachability TLV defined in this document:

Sub-TLV type	Length	Name
24	4	remote AS number
25	4	IPv4 remote ASBR identifier
26	16	IPv6 remote ASBR identifier
TBD1	16	IPv6 Router ID

Detailed definitions of the three new sub-TLVs are described in Section 3.3.1, 3.3.2, 3.3.3, and 3.3.4.

3.2. TE Router ID

The IPv4 TE Router ID TLV and IPv6 TE Router ID TLV, which are defined in [RFC5305] and [RFC6119] respectively, only have area flooding-scope. When performing inter-AS TE, the TE Router ID MAY be needed to reach all routers within an entire ISIS routing domain and it MUST have the same flooding scope as the Inter-AS Reachability TLV does.

[RFC4971] defines a generic advertisement mechanism for ISIS which allows a router to advertise its capabilities within an ISIS area or an entire ISIS routing domain. [RFC4971] also points out that the TE

Router ID is a candidate to be carried in the IS-IS router capability TLV when performing inter-area TE.

This document uses such mechanism for TE Router ID advertisement when the TE Router ID is needed to reach all routers within an entire ISIS Routing domain. Two new sub-TLVs are defined for inclusion in the IS-IS Router Capability TLV to carry the TE Router IDs.

Sub-TLV type	Length	Name
11	4	IPv4 TE Router ID
12	16	IPv6 TE Router ID

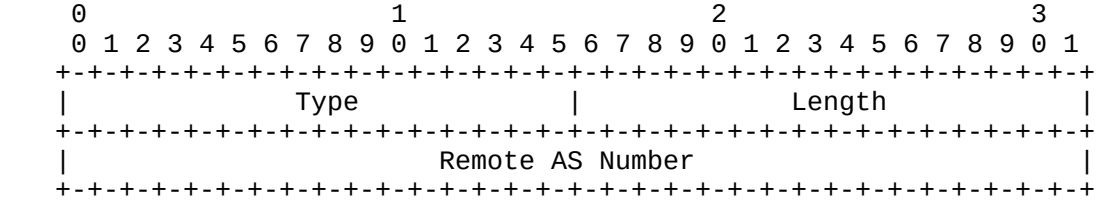
Detailed definitions of the new sub-TLV are described in Section 3.4.1 and 3.4.2.

3.3. Sub-TLVs for Inter-AS Reachability TLV

3.3.1. Remote AS Number Sub-TLV

A new sub-TLV, the remote AS number sub-TLV, is defined for inclusion in the inter-AS reachability TLV when advertising inter-AS links. The remote AS number sub-TLV specifies the AS number of the neighboring AS to which the advertised link connects.

The remote AS number sub-TLV is TLV type 24 (see Section 6.2) and is 4 octets in length. The format is as follows:



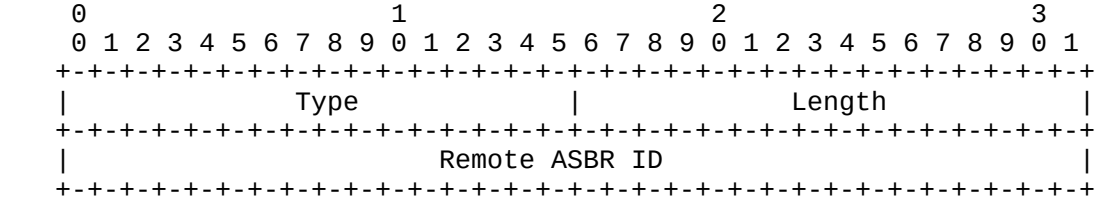
The remote AS number field has 4 octets. When only 2 octets are used for the AS number, as in current deployments, the left (high-order) 2 octets MUST be set to 0. The remote AS number sub-TLV MUST be included when a router advertises an inter-AS TE link.

3.3.2. IPv4 Remote ASBR ID Sub-TLV

A new sub-TLV, which is referred to as the IPv4 remote ASBR ID sub-TLV, is defined for inclusion in the inter-AS reachability TLV when advertising inter-AS links. The IPv4 remote ASBR ID sub-TLV specifies the IPv4 identifier of the remote ASBR to which the advertised inter-AS link connects. This could be any stable and

routable IPv4 address of the remote ASBR. Use of the TE Router ID as specified in the Traffic Engineering router ID TLV [RFC5305] is RECOMMENDED.

The IPv4 remote ASBR ID sub-TLV is TLV type 25 (see Section 6.2) and is 4 octets in length. The format of the IPv4 remote ASBR ID sub-TLV is as follows:

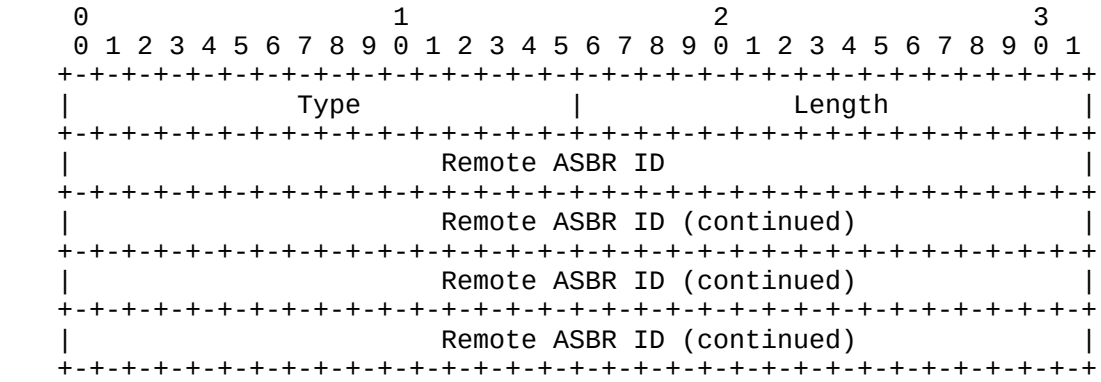


The IPv4 remote ASBR ID sub-TLV MUST be included if the neighboring ASBR has an IPv4 address. If the neighboring ASBR does not have an IPv4 address (not even an IPv4 TE Router ID), the IPv6 remote ASBR ID sub-TLV MUST be included instead. An IPv4 remote ASBR ID sub-TLV and IPv6 remote ASBR ID sub-TLV MAY both be present in an extended IS reachability TLV.

3.3.3. IPv6 Remote ASBR ID Sub-TLV

A new sub-TLV, which is referred to as the IPv6 remote ASBR ID sub-TLV, is defined for inclusion in the inter-AS reachability TLV when advertising inter-AS links. The IPv6 remote ASBR ID sub-TLV specifies the IPv6 identifier of the remote ASBR to which the advertised inter-AS link connects. This could be any stable and routable IPv6 address of the remote ASBR. Use of the TE Router ID as specified in the IPv6 Traffic Engineering router ID TLV [RFC6119] is RECOMMENDED.

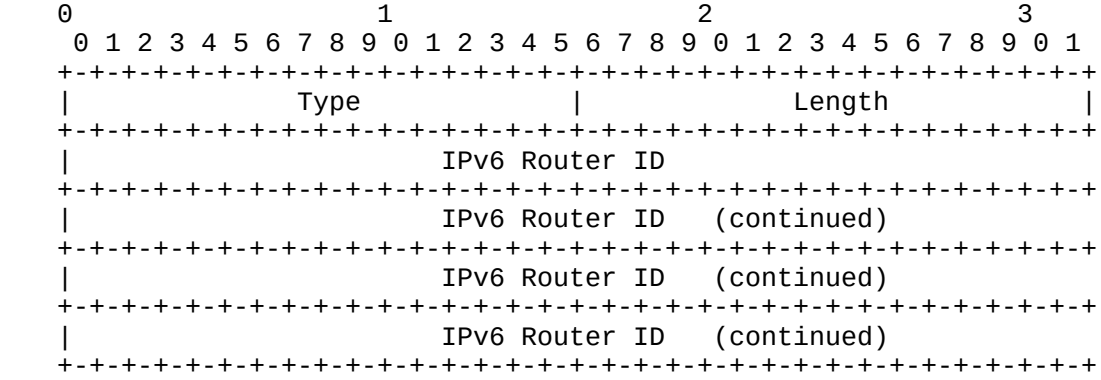
The IPv6 remote ASBR ID sub-TLV is TLV type 26 (see Section 6.2) and is 16 octets in length. The format of the IPv6 remote ASBR ID sub-TLV is as follows:



The IPv6 remote ASBR ID sub-TLV MUST be included if the neighboring ASBR has an IPv6 address. If the neighboring ASBR does not have an IPv6 address, the IPv4 remote ASBR ID sub-TLV MUST be included instead. An IPv4 remote ASBR ID sub-TLV and IPv6 remote ASBR ID sub-TLV MAY both be present in an extended IS reachability TLV.

3.3.4. IPv6 Router ID sub-TLV

The IPv6 Router ID sub-TLV is TLV type TBD1 (see Section 6.3) and is 16 octets in length. The format of the IPv6 Router ID sub-TLV is as follows:



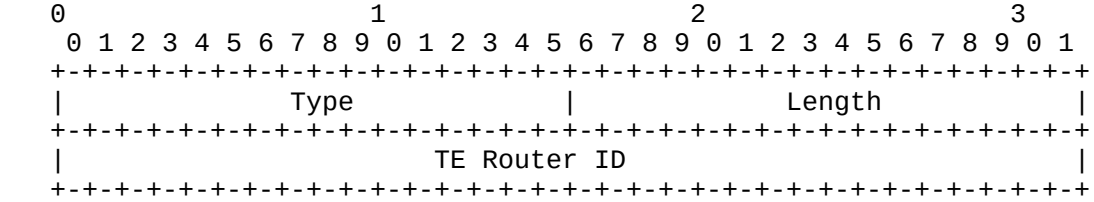
The IPv6 TE Router ID SHOULD be identical to the value advertised in the IPv6 Traffic Engineering Router ID TLV [RFC6119].

If the originating node does not support IPv4, the IPv6 Router ID sub-TLV MUST be present in the inter-AS reachability TLV. Inter-AS reachability TLVs which have a Router ID of 0.0.0.0 and do NOT have the IPv6 Router ID sub-TLV present MUST be ignored.

3.4. Sub-TLVs for IS-IS Router Capability TLV

3.4.1. IPv4 TE Router ID sub-TLV

The IPv4 TE Router ID sub-TLV is TLV type 11 (see Section 6.3) and is 4 octets in length. The format of the IPv4 TE Router ID sub-TLV is as follows:

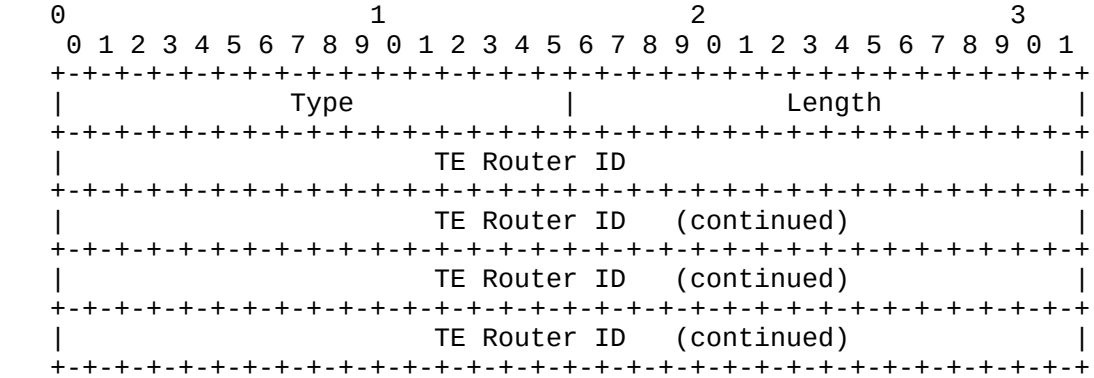


The IPv4 TE Router ID SHOULD be identical to the value advertised in the IPv4 Traffic Engineering Router ID TLV [RFC5305].

When the TE Router ID is needed to reach all routers within an entire ISIS routing domain, the IS-IS Router capability TLV MUST be included in its LSP. If an ASBR supports Traffic Engineering for IPv4 and if the ASBR has an IPv4 TE Router ID, the IPv4 TE Router ID sub-TLV MUST be included. If the ASBR does not have an IPv4 TE Router ID, the IPv6 TE Router sub-TLV MUST be included instead. An IPv4 TE Router ID sub-TLV and IPv6 TE Router ID sub-TLV MAY both be present in an IS-IS router capability TLV.

3.4.2. IPv6 TE Router ID sub-TLV

The IPv6 TE Router ID sub-TLV is TLV type 12 (see Section 6.3) and is 16 octets in length. The format of the IPv6 TE Router ID sub-TLV is as follows:



The IPv6 TE Router ID SHOULD be identical to the value advertised in the IPv6 Traffic Engineering Router ID TLV [RFC6119].

When the TE Router ID is needed to reach all routers within an entire ISIS routing domain, the IS-IS router capability TLV MUST be included in its LSP. If an ASBR supports Traffic Engineering for IPv6 and if the ASBR has an IPv6 TE Router ID, the IPv6 TE Router ID sub-TLV MUST be included. If the ASBR does not have an IPv6 TE Router ID, the IPv4 TE Router sub-TLV MUST be included instead. An IPv4 TE Router ID sub-TLV and IPv6 TE Router ID sub-TLV MAY both be present in an IS-IS router capability TLV.

4. Procedure for Inter-AS TE Links

When TE is enabled on an inter-AS link and the link is up, the ASBR SHOULD advertise this link using the normal procedures for [RFC5305]. When either the link is down or TE is disabled on the link, the ASBR SHOULD withdraw the advertisement. When there are changes to the TE parameters for the link (for example, when the available bandwidth changes), the ASBR SHOULD re-advertise the link but MUST take precautions against excessive re-advertisements.

Hellos MUST NOT be exchanged over the inter-AS link, and consequently, an ISIS adjacency MUST NOT be formed.

The information advertised comes from the ASBR’s knowledge of the TE capabilities of the link, the ASBR’s knowledge of the current status and usage of the link, and configuration at the ASBR of the remote AS number and remote ASBR TE Router ID.

Legacy routers receiving an advertisement for an inter-AS TE link are able to ignore it because they do not know the new TLV and sub-TLVs that are defined in Section 3 of this document. They will continue to flood the LSP, but will not attempt to use the information received.

In the current operation of ISIS TE, the LSRs at each end of a TE link emit LSPs describing the link. The databases in the LSRs then have two entries (one locally generated, the other from the peer) that describe the different ’directions’ of the link. This enables Constrained Shortest Path First (CSPF) to do a two-way check on the link when performing path computation and eliminate it from consideration unless both directions of the link satisfy the required constraints.

In the case we are considering here (i.e., of a TE link to another AS), there is, by definition, no IGP peering and hence no bidirectional TE link information. In order for the CSPF route

computation entity to include the link as a candidate path, we have to find a way to get LSPs describing its (bidirectional) TE properties into the TE database.

This is achieved by the ASBR advertising, internally to its AS, information about both directions of the TE link to the next AS. The ASBR will normally generate a LSP describing its own side of a link; here we have it 'proxy' for the ASBR at the edge of the other AS and generate an additional LSP that describes that device's 'view' of the link.

Only some essential TE information for the link needs to be advertised; i.e., the Interface Address, the remote AS number, and the remote ASBR ID of an inter-AS TE link.

Routers or PCEs that are capable of processing advertisements of inter-AS TE links SHOULD NOT use such links to compute paths that exit an AS to a remote ASBR and then immediately re-enter the AS through another TE link. Such paths would constitute extremely rare occurrences and SHOULD NOT be allowed except as the result of specific policy configurations at the router or PCE computing the path.

4.1. Origin of Proxied TE Information

Section 4 describes how an ASBR advertises TE link information as a proxy for its neighbor ASBR, but does not describe where this information comes from.

Although the source of this information is outside the scope of this document, it is possible that it will be a configuration requirement at the ASBR, as are other local properties of the TE link. Further, where BGP is used to exchange IP routing information between the ASBRs, a certain amount of additional local configuration about the link and the remote ASBR is likely to be available.

We note further that it is possible, and may be operationally advantageous, to obtain some of the required configuration information from BGP. Whether and how to utilize these possibilities is an implementation matter.

5. Security Considerations

The protocol extensions defined in this document are relatively minor and can be secured within the AS in which they are used by the existing ISIS security mechanisms (e.g., using the cleartext passwords or Hashed Message Authentication Codes - Message Digest 5

(HMAC-MD5) algorithm, which are defined in [RFC1195] and [RFC3567] separately).

There is no exchange of information between ASes, and no change to the ISIS security relationship between the ASes. In particular, since no ISIS adjacency is formed on the inter-AS links, there is no requirement for ISIS security between the ASes.

Some of the information included in these new advertisements (e.g., the remote AS number and the remote ASBR ID) is obtained manually from a neighboring administration as part of a commercial relationship. The source and content of this information should be carefully checked before it is entered as configuration information at the ASBR responsible for advertising the inter-AS TE links.

It is worth noting that in the scenario we are considering, a Border Gateway Protocol (BGP) peering may exist between the two ASBRs and that this could be used to detect inconsistencies in configuration (e.g., the administration that originally supplied the information may be lying, or some manual mis-configurations or mistakes may be made by the operators). For example, if a different remote AS number is received in a BGP OPEN [RFC4271] from that locally configured to ISIS-TE, as we describe here, then local policy SHOULD be applied to determine whether to alert the operator to a potential mis-configuration or to suppress the ISIS advertisement of the inter-AS TE link. Note further that if BGP is used to exchange TE information as described in Section 4.1, the inter-AS BGP session SHOULD be secured using mechanisms as described in [RFC4271] to provide authentication and integrity checks.

For a discussion of general security considerations for IS-IS, see [RFC5304].

6. IANA Considerations

IANA is requested to make the following allocations from registries under its control.

6.1. Inter-AS Reachability TLV

This document defines the following new ISIS TLV type, described in Section 3.1, which has been registered in the ISIS TLV codepoint registry:

Type	Description	IIH	LSP	SNP
141	inter-AS reachability information	n	y	n

6.2. Sub-TLVs for the Inter-AS Reachability TLV

This document defines the following new sub-TLV types (described in Sections 3.3.1, 3.3.2, 3.3.3, and, 3.3.4) of top-level TLV 141 (see Section 6.1 above), which have been registered in the ISIS sub-TLV registry for TLV 141. Note that these four new sub-TLVs SHOULD NOT appear in TLV 22 (or TLV 23, TLV 222, TLV223) and MUST be ignored in TLV 22 (or TLV 23, TLV 222, TLV223):

Type	Description
24	remote AS number
25	IPv4 remote ASBR identifier
26	IPv6 remote ASBR identifier
TBD1	IPv6 Router ID

As described above in Section 3.1, the sub-TLVs which are defined in [RFC5305], [RFC6119] and other documents for describing the TE properties of an TE link are applicable to describe an inter-AS TE link and MAY be included in the inter-AS reachability TLV when adverting inter-AS TE links.

IANA has created the following sub-TLVs registries in "Sub-TLVs for TLVs 22, 23, 141, 222, and 223" registry.

Type	Description	TLV 22	TLV 23	TLV 141	TLV 222	TLV 223	Reference
24	remote AS number	n	n	y	n	n	[This.I-D]
25	IPv4 remote ASBR identifier	n	n	y	n	n	[This.I-D]
26	IPv6 remote ASBR identifier	n	n	y	n	n	[This.I-D]

IANA is requested to create a new sub-TLV registry in "Sub-TLVs for TLVs 22, 23, 141, 222, and 223" registry.

Type	Description	TLV 22	TLV 23	TLV 141	TLV 222	TLV 223	Reference
TBD1	IPv6 Router ID	n	n	y	n	n	[This.I-D]

6.3. Sub-TLVs for the IS-IS Router Capability TLV

This document defines the following new sub-TLV types, described in Sections 3.4.1 and 3.4.2, of top-level TLV 242 (which is defined in [RFC4971]) that have been registered in the ISIS sub-TLV registry for TLV 242:

Type	Description	Length
11	IPv4 TE Router ID	4
12	IPv6 TE Router ID	16

7. Acknowledgements

For the original version of [RFC5316] the authors would like to thank Adrian Farrel, Jean-Louis Le Roux, Christian Hopps, Les Ginsberg, and Hannes Gredler for their review and comments on this document.

8. References

8.1. Normative References

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <http://www.rfc-editor.org/info/rfc1195>.

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

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

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <http://www.rfc-editor.org/info/rfc6119>.

8.2. Informative References

[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, <http://www.rfc-editor.org/info/rfc3209>.

[RFC3567] Li, T. and R. Atkinson, "Intermediate System to Intermediate System (IS-IS) Cryptographic Authentication", RFC 3567, DOI 10.17487/RFC3567, July 2003, <http://www.rfc-editor.org/info/rfc3567>.

[RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, DOI 10.17487/RFC4206, October 2005, <http://www.rfc-editor.org/info/rfc4206>.

[RFC4216] Zhang, R., Ed. and J. Vasseur, Ed., "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, DOI 10.17487/RFC4216, November 2005, <http://www.rfc-editor.org/info/rfc4216>.

[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, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <http://www.rfc-editor.org/info/rfc4655>.

[RFC4971] Vasseur, JP., Ed., Shen, N., Ed., and R. Aggarwal, Ed., "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, DOI 10.17487/RFC4971, July 2007, <http://www.rfc-editor.org/info/rfc4971>.

[RFC5152] Vasseur, JP., Ed., Ayyangar, A., Ed., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, DOI 10.17487/RFC5152, February 2008, <http://www.rfc-editor.org/info/rfc5152>.

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

[RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <http://www.rfc-editor.org/info/rfc5307>.

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

[RFC5441] Vasseur, JP., Ed., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, DOI 10.17487/RFC5441, April 2009, <http://www.rfc-editor.org/info/rfc5441>.

[RFC6823] Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, DOI 10.17487/RFC6823, December 2012, <http://www.rfc-editor.org/info/rfc6823>.

Appendix A. Changes to RFC 5316

This document makes the following changes to RFC 5316.

RFC 5316 only allowed a 32 bit Router ID in the fixed header of TLV 141. This is problematic in an IPv6-only deployment where an IPv4 address may not be available. This document specifies:

1. The Router ID SHOULD be identical to the value advertised in the Traffic Engineering Router ID TLV (134) if available.

2. If no Traffic Engineering Router ID is assigned the Router ID SHOULD be identical to an IP Interface Address [RFC1195] advertised by the originating IS.

3. If the originating node does not support IPv4, then the reserved value 0.0.0.0 MUST be used in the Router ID field and the IPv6 TE Router ID sub-TLV MUST be present in the TLV.

Authors' Addresses

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com

Les Ginsberg
Cisco Systems

Email: ginsberg@cisco.com

Stefano Previdi
Cisco Systems

Email: sprevidi@cisco.com

Xiaodong Duan
China Mobile

Email: duanxiaodong@chinamobile.com

isis
Internet-Draft
Intended status: Standards Track
Expires: November 10, 2017

B. Liu, Ed.
Huawei Technologies
L. Ginsberg
Cisco Systems
B. Decraene
Orange
I. Farrer
Deutsche Telekom AG
M. Abrahamsson
T-Systems

May 9, 2017

ISIS Auto-Configuration
draft-ietf-isis-auto-conf-05

Abstract

This document specifies IS-IS auto-configuration mechanisms. The key components are IS-IS System ID self-generation, duplication detection and duplication resolution. These mechanisms provide limited IS-IS functions, and so are suitable for networks where plug-and-play configuration is expected.

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 [RFC2119] when they appear in ALL CAPS. When these words are not in ALL CAPS (such as "should" or "Should"), they have their usual English meanings, and are not to be interpreted as [RFC2119] key words.

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 http://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 November 10, 2017.

Table of Contents

1. Introduction

This document specifies mechanisms for IS-IS [RFC1195] [ISO_IEC10589][RFC5308] to be auto-configuring. Such mechanisms could reduce the management burden for configuring a network, especially where plug-and-play device configuration is required.

IS-IS auto-configuration is comprised of the following functions:

1. IS-IS default configuration.

2. IS-IS System ID self-generation.

3. System ID duplication detection and resolution.

4. ISIS TLV utilization (Authentication TLV, metrics in reachability advertisements, and Dynamic Host Name TLV).

This document also defines mechanisms to prevent the unintentional interoperation of auto-configured routers with non-autoconfigured routers. See Section 3.3.

2. Scope

The auto-configuration mechanisms support both IPv4 and IPv6 deployments.

These auto-configuration mechanisms aim to cover simple deployment cases. The following important features are not supported:

- Multiple IS-IS instances.

- Multi-area and level-2 routing.

- Interworking with other routing protocols.

IS-IS auto-configuration is primarily intended for use in small (i.e. 10s of devices) and unmanaged deployments. It allows IS-IS to be used without the need for any configuration by the user. It is not recommended for larger deployments.

3. Protocol Specification

3.1. IS-IS Default Configuration

- IS-IS interfaces MUST be auto-configured to an interface type corresponding to their layer-2 capability. For example, Ethernet interfaces will be auto-configured as broadcast networks and

Point-to-Point Protocol (PPP) interfaces will be auto-configured as Point-to-Point interfaces.

- IS-IS auto-configuration instances MUST be configured as level-1, so that the interfaces operate as level-1 only.
- originatingLSPBufferSize is set to 512.
- MaxAreaAddresses is set to 3
- Extended IS Reachability and IP Reachability TLVs [RFC5305] MUST be used i.e. a router operating in auto configuration mode MUST NOT use any of the following TLVs:
  * IS Neighbors (2)
  * IP Internal Reachability (128)
  * IP External Reachability (130)

  TLVs listed above MUST be ignored on receipt.

3.2. IS-IS NET Generation

In IS-IS, a router (known as an Intermediate System) is identified by a Network Entity Title (NET) which is a type of Network Service Access Point (NSAP). The NET is the address of an instance of the IS-IS protocol running on an Intermediate System (IS).

The auto-configuration mechanism generates the IS-IS NET as the following:

- Area address

  In IS-IS auto-configuration, this field MUST be 13 octets long and set to all 0.

- System ID

  This field follows the area address field, and is 6 octets in length. There are two basic requirements for the System ID generation:

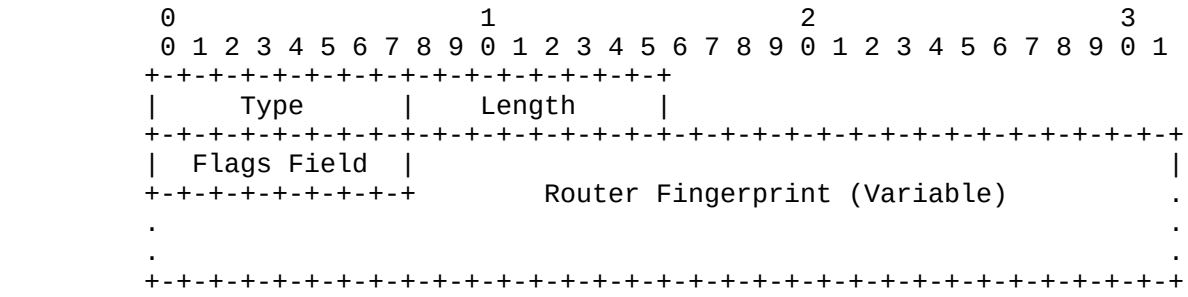
  - As specified by the IS-IS protocol, this field must be unique among all routers in the same area.
  - After its initial generation, the System ID SHOULD remain stable. Changes such as interface enable/disable, interface

connect/disconnect, device reboot, firmware update, or configuration changes SHOULD NOT cause the system ID to change. System ID change as part of the System ID collision resolution process MUST be supported. Implementations SHOULD allow the System ID to be cleared by a user initiated system reset.

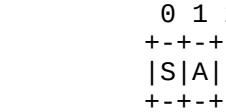
More specific considerations for System ID generation are described in Section 3.4.5.

3.3. Router-Fingerprint TLV

The Router-Fingerprint TLV is similar to the Router-Hardware-Fingerprint TLV defined in [RFC7503]. However, the TLV defined here includes a flags field to support indicating that the router is in Start-up mode and is operating in auto-configuration mode.



Type: to be assigned by IANA.
Length: the length of the value field. Must be >= 33.
Flags field (1 octet)



S flag: when set, indicates the router is in "start-up" mode.
A flag: when set, indicates that the router is operating in auto-configuration mode. The purpose of the flag is so that two routers can identify if they are both using auto-configuration. If the A flag setting does not match in hellos then no adjacency should be formed.
Reserved: these bits MUST be set to zero and MUST be ignored by the receiver.

Router Fingerprint: 32 or more octets.

More specific considerations for Router-Fingerprint are described in Section 3.4.5.

Router Fingerprint TLV MUST be included in Intermediate System to Intermediate System Hellos (IIHs) originated by a router operating in auto-configuration mode. An auto-configuration mode router MUST ignore IIHs that don't contain the Router Fingerprint TLV.

Router Fingerprint TLV MUST be included in Link State PDU (LSP) #0 originated by a router operating in auto-configuration mode. If an LSP #0 which does NOT contain a Router Fingerprint TLV is received by a Router operating in auto-configuration mode the LSP is flooded as normal, but the entire LSP set originated by the sending router MUST be ignored when running the Decision process.

The router fingerprint TLV MUST NOT be included in an LSP with a non-zero number and when received MUST be ignored.

3.4. Protocol Operation

This section describes the operation of a router supporting auto-configuration mode.

3.4.1. Start-Up mode

When a router starts operation in auto-configuration mode, both the S and A bits MUST be set in the Router Fingerprint TLV included in both hellos and LSP #0. During this mode only LSP #0 is generated and IS or IP/IPv6 reachability TLVs MUST NOT be included in LSP #0. A router remains in Start-up mode for a minimum period of time (recommended to be 1 minute). This time should be sufficient to bring up adjacencies to all expected neighbors. A router leaves Start-up mode once the minimum time has elapsed and full LSP database synchronization is achieved with all neighbors in the UP state.

When a router exits startup-mode it clears the S bit in Router Fingerprint TLVs it sends in hellos and LSP#0. The router MAY now advertise IS neighbor and IP/IPv6 prefix reachability in its LSPs and MAY generate LSPs with a non-zero number.

The purpose of Start-up Mode is to minimize the occurrence of System ID changes for a router once it has become fully operational. Any System ID change during Start-up mode will have minimal impact on a running network because while in Start-up mode the router is not yet being used for forwarding traffic.

3.4.2. Adjacency Formation

Routers operating in auto-configuration mode MUST NOT form adjacencies with routers which are NOT operating in auto-configuration mode. The presence of the Router Fingerprint TLV with the A bit set indicates the router is operating in auto-configuration mode.

NOTE: The use of the special area address of all 0's makes it unlikely that a router which is not operating in auto-configuration mode will be in the same area as a router operating in auto-configuration mode. However, the check for the Router Fingerprint TLV with A bit set provides additional protection.

3.4.3. IS-IS System ID Duplication Detection

The System ID of each node MUST be unique. As described in Section 3.4.5, the System ID is generated based on entropies (e.g. MAC address) which are generally expected to be unique. However, since there may be limitations to the available entropies, there is still the possibility of System ID duplication. This section defines how IS-IS detects and resolves System ID duplication. Duplicate System ID may occur between neighbors or between routers in the same area which are not neighbors.

Duplicate System ID with a neighbor is detected when the System ID received in an IIH is identical to the local System ID and the Router-Fingerprint in the received Router-Fingerprint TLV does NOT match the locally generated Router-Fingerprint.

Duplicate System ID with a non-neighbor is detected when an LSP #0 is received, the System ID of the originator is identical to the local System ID, and the Router-Fingerprint in the Router-Fingerprint TLV does NOT match the locally generated Router-Fingerprint.

3.4.4. Duplicate System ID Resolution Procedures

When duplicate System ID is detected one of the systems MUST assign itself a different System ID and perform a protocol restart. The resolution procedure attempts to minimize disruption to a running network by choosing a router which is in Start-up mode to be restarted whenever possible.

The contents of the Router-Fingerprint TLVs for the two routers with duplicate System IDs are compared.

If one TLV has the S bit set (router is in Start-up mode) and one TLV has the S bit clear (router is NOT in Start-up mode) the router in Start-up mode MUST generate a new System ID and restart the protocol.

If both TLVs have the S bit set (both routers are in Start-up mode) or both TLVs have the S bit clear (neither router is in Start-up mode) then the router with numerically smaller Router-Fingerprint MUST generate a new System ID and restart the protocol.

Fingerprint comparison is performed octet by octet starting from the first received octet until a difference is detected. If the fingerprints have different lengths and all octets up to the shortest length are identical then the fingerprint with smaller length is considered smaller.

If the fingerprints are identical in both content and length (and state of the S bit is identical) and the duplication is detected in hellos then the both routers MUST generate a new System ID and restart the protocol.

If fingerprints are identical in both content and length and the duplication is detected in LSP #0 then the procedures defined in Section 3.4.6 MUST be followed.

3.4.5. System ID and Router-Fingerprint Generation Considerations

As specified in this document, there are two distinguishing items that need to be self-generated: the System ID and Router-Fingerprint. In a network device, normally there are some resources which can provide an extremely high probability of uniqueness (some examples listed below). These resources can be used as seeds to derive identifiers.

- MAC address(es)
- Configured IP address(es)
- Hardware IDs (e.g. CPU ID)
- Device serial number(s)
- System clock at a certain specific time
- Arbitrary received packet(s) on an interface(s)

This document recommends the use of an IEEE 802 48-bit MAC address associated with the router as the initial System ID. This document

does not specify a specific method to re-generate the System ID when duplication happens.

This document also does not specify a specific method to generate the Router-Fingerprint.

There is an important concern that the seeds listed above (except MAC address) might not be available in some small devices such as home routers. This is because of hardware/software limitations and the lack of sufficient communication packets at the initial stage in home routers when doing ISIS auto-configuration. In this case, this document suggests using the MAC address as System ID and generating a pseudo-random number based on another seed (such as the memory address of a certain variable in the program) as the Router-Fingerprint. The pseudo-random number might not have a very high probability of uniqueness in this solution, but should be sufficient in home networks scenarios.

The considerations surrounding System ID stability described in section Section 3.2 also need to be applied.

3.4.6. Duplication of both System ID and Router-Fingerprint

As described above, the resources for generating System ID/ Fingerprint might be very constrained during the initial stages. Hence, the duplication of both System ID and Router-Fingerprint needs to be considered. In such a case it is possible that a router will receive an LSP with System ID and Router-Fingerprint identical to the local values but the LSP is NOT identical to the locally generated copy i.e. sequence number is newer or sequence number is the same but the LSP has a valid checksum which does not match. The term DD-LSP is used to describe such an LSP.

In a benign case, this will occur if a router restarts and it receives copies of its own LSPs from its previous incarnation. This benign case needs to be distinguished from the pathological case where there are two different routers with the same System ID and the same Router-Fingerprint.

In the benign case, the restarting router will generate a new version of its own LSP with higher sequence number and flood the new LSP version. This will cause other routers in the network to update their LSPDB and synchronization will be achieved.

In the pathological case the generation of a new version of an LSP by one of the "twins" will cause the other twin to generate the same LSP with a higher sequence number - and oscillation will continue without achieving LSPDB synchronization.

Note that comparison of S bit in the Router-Fingerprint TLV cannot be performed as in the benign case it is expected that the S bit will be clear. Also note that the conditions for detecting duplicate System ID will NOT be satisfied because both the System ID and the Router-Fingerprint will be identical.

The following procedure is defined:

```
    DD-state is a boolean which indicates if a
      DD-LSP #0 has been received
    DD-count is the count of the number of occurences
      of reception of a DD-LSP
    DD-timer is a timer associated with reception of
     DD-LSPs. Recommended value is 60 seconds.
    DD-max is the maximum number of DD-LSPs allowed
     to be received in DD-timer interval.
     Recommended value is 3.
```

When a DD-LSP is received:

```
  If DD-state is FALSE:
    DD-state is set to TRUE
    DD-timer is started
    DD-count is initialized to 1.

  If DD-state is TRUE:
    DD-count is incremented
    If DD-count is >= DD-max:
       Local system MUST generate a new System ID
        and Router-Fingerprint and restart the protocol
       DD-state is (re)initialized to FALSE and
        DD-timer cancelled.

  If DD-timer expires:
    DD-state is set to FALSE.
```

Note that to minimze the likelihood of duplication of both System ID and Router-fingerprint reoccuring, routers SHOULD have more entropies available. One simple way to achieve this is to add the LSP sequence number of the next LSP it will send to the Router-Fingerprint.

3.5. Additional IS-IS TLVs Usage Guidelines

This section describes the behavior of selected TLVs when used by a router supporting IS-IS auto-configuration.

3.5.1. Authentication TLV

It is RECOMMENDED that IS-IS routers supporting this specification offer an option to explicitly configure a single password for HMAC-MD5 authentication as specified in[RFC5304].

3.5.2. Metric Used in Reachability TLVs

It is RECOMMENDED that IS-IS auto-configuration routers use a high metric value (e.g. 100000) as default in order to allow manually configured adjacencies to be preferred over auto-configured.

3.5.3. Dynamic Host Name TLV

IS-IS auto-configuration routers MAY advertise their Dynamic Host Name TLV (TLV 137, [RFC5301]). The host name could be provisioned by an IT system, or just use the name of vendor, device type or serial number, etc.

To guarantee the uniqueness of the host name, the System ID SHOULD be appended as a suffix in the names.

4. Security Considerations

In the absence of cryptographic authentication it is possible for an attacker to inject a PDU falsely indicating there is a duplicate system-id. This may trigger automatic restart of the protocol using the duplicate-id resolution procedures defined in this document.

Note that the use of authentication is incompatible with auto-configuration as it requires some manual configuration.

For wired deployment, the wired connection itself could be considered as an implicit authentication in that unwanted routers are usually not able to connect (i.e. there is some kind of physical security in place preventing the connection of rogue devices); for wireless deployment, the authentication could be achieved at the lower wireless link layer.

5. IANA Considerations

This document requires the definition of a new IS-IS TLV to be reflected in the "IS-IS TLV Codepoints" registry:

Type	Description	IIH	LSP	SNP	Purge
TBA	Router-Fingerprint	Y	Y	N	Y

6. Acknowledgements

This document was heavily inspired by [RFC7503].

Martin Winter, Christian Franke and David Lamparter gave essential feedback to improve the technical design based on their implementation experience.

Many useful comments were made by Acee Lindem, Karsten Thomann, Hannes Gredler, Peter Lothberg, Uma Chundury, Qin Wu, Sheng Jiang and Nan Wu, etc.

This document was produced using the xml2rfc tool [RFC7991]. (initially prepared using 2-Word-v2.0.template.dot.)

7. References

7.1. Normative References

[ISO_IEC10589] "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <http://www.rfc-editor.org/info/rfc1195>.

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

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

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

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

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

7.2. Informative References

[RFC7503] Lindem, A. and J. Arkko, "OSPFv3 Autoconfiguration", RFC 7503, DOI 10.17487/RFC7503, April 2015, <http://www.rfc-editor.org/info/rfc7503>.

[RFC7991] Hoffman, P., "The "xml2rfc" Version 3 Vocabulary", RFC 7991, DOI 10.17487/RFC7991, December 2016, <http://www.rfc-editor.org/info/rfc7991>.

Authors' Addresses

Bing Liu (editor)
Huawei Technologies
Q10, Huawei Campus, No.156 Beiqing Road
Hai-Dian District, Beijing, 100095
P.R. China

Email: leo.liubing@huawei.com

Les Ginsberg
Cisco Systems
821 Alder Drive
Milpitas CA 95035
USA

Email: ginsberg@cisco.com

Bruno Decraene
Orange
France

Email: bruno.decraene@orange.com

Ian Farrer
Deutsche Telekom AG
Bonn
Germany

Email: ian.farrer@telekom.de

Mikael Abrahamsson
T-Systems
Stockholm
Sweden

Email: mikael.abrahamsson@t-systems.se

Networking Working Group
Internet-Draft
Obsoletes: 6822 (if approved)
Intended status: Standards Track
Expires: October 19, 2017

L. Ginsberg
S. Previdi
Cisco Systems
W. Henderickx
Nokia
April 17, 2017

IS-IS Multi-Instance
draft-ietf-isis-mi-bis-03.txt

Abstract

This document describes a mechanism that allows a single router to share one or more circuits among multiple Intermediate System To Intermediate System (IS-IS) routing protocol instances.

Multiple instances allow the isolation of resources associated with each instance. Routers will form instance specific adjacencies. Each instance can support multiple topologies. Each topology has a unique Link State Database (LSDB). Each Protocol Data Unit (PDU) will contain a new Type-Length-Value (TLV) identifying the instance and the topology (or topologies) to which the PDU belongs.

This document obsoletes RFC 6822 if approved.

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

Table of Contents

1. Introduction

An existing limitation of the protocol defined by [ISO10589] is that only one instance of the protocol can operate on a given circuit. This document defines an extension to IS-IS to remove this restriction. The extension is referred to as "Multi-instance IS-IS" (MI-IS-IS).

Routers that support this extension are referred to as "Multi-Instance-capable routers" (MI-RTR).

The use of multiple instances enhances the ability to isolate the resources associated with a given instance both within a router and across the network. Instance-specific prioritization for processing PDUs and performing routing calculations within a router may be specified. Instance-specific flooding parameters may also be defined so as to allow different instances to consume network-wide resources at different rates.

Another existing protocol limitation is that a given instance supports a single Update Process operating on a single Link State Database (LSDB). This document defines an extension to IS-IS to allow non-zero instances of the protocol to support multiple Update Processes. Each Update Process is associated with a topology and a unique topology specific LSDB. Non-zero instances of the protocol are only supported by MI-RTRs. Legacy routers support the standard or zero instance of the protocol. The behavior of the standard instance is not changed in any way by the extensions defined in this document.

MI-IS-IS might be used to support topology-specific routing. Two methods of supporting such a use are defined in this document. One supports the use of [RFC5120] within a reserved instance specific topology. The second is an alternative to [RFC5120] which supports topology-specific flooding of link state information.

MI-IS-IS might also be used to support advertisement of information on behalf of applications [RFC6823]. The advertisement of information not directly related to the operation of the IS-IS protocol can therefore be done in a manner that minimizes its impact on the operation of routing.

The above are examples of how MI-IS-IS might be used. The specification of uses of MI-IS-IS is outside the scope of this document.

2. Elements Of Procedure

An Instance Identifier (IID) is introduced to uniquely identify an IS-IS instance. The protocol extension includes a new TLV (IID-TLV) in each IS-IS PDU originated by an MI-RTR except as noted in this document. The IID-TLV identifies the unique instance as well as the instance-specific topology/topologies to which the PDU applies. Each IS-IS PDU is associated with only one IS-IS instance.

MI-RTRs form instance-specific adjacencies. The IID-TLV included in IS-IS Hellos (IIH) includes the IID and the set of Instance-Specific Topology Identifiers (ITIDs) that the sending IS supports. When multiple instances share the same circuit, each instance will have a separate set of adjacencies.

MI-RTRs support the exchange of topology-specific Link State PDUs for the IID/ITID pairs that each neighbor supports. A unique IS-IS Update Process (see [ISO10589] operates for each IID/ITID pair. This MAY also imply IID/ITID-specific routing calculations and IID/ITID-specific routing and forwarding tables. However, this aspect is outside the scope of this specification.

The mechanisms used to implement support of the separation of IS-IS instances and topology-specific Update Processes within a router are outside the scope of this specification.

2.1. Instance Identifier TLV

A new TLV is defined in order to convey the IID and ITIDs supported. The IID-TLV associates a PDU with an IS-IS instance using a unique 16-bit number. The IID-TLV is carried in all IS-IS PDUs that are associated with a non-zero instance; this includes IIHs, Sequence Number PDUs (SNPs), and Link State PDUs (LSPs) .

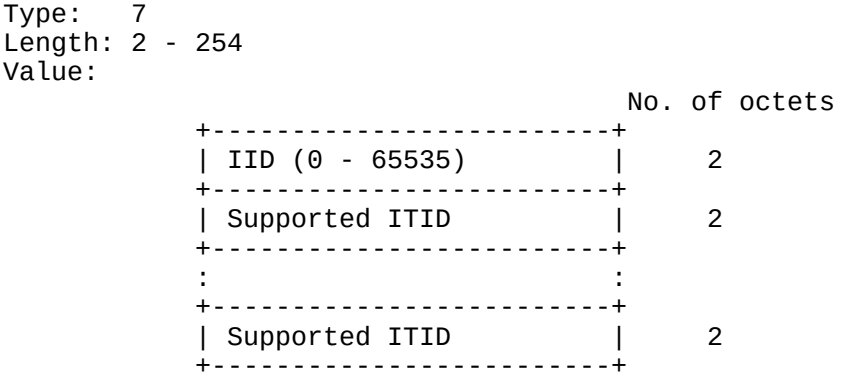
Multiple instances of IS-IS may coexist on the same circuit and on the same physical router. IIDs MUST be unique within the same routing domain.

IID #0 is reserved for the standard instance supported by legacy systems. IS-IS PDUs associated with the standard instance MUST NOT include an IID-TLV except where noted in this document.

The IID-TLV MAY include one or more ITIDs. An ITID is a 16-bit identifier where all values (0 - 65535) are valid.

The following format is used for the IID-TLV:

Type: 7
Length: 2 - 254
Value:



When the IID = 0, the list of supported ITIDs MUST NOT be present.

An IID-TLV with IID = 0 MUST NOT appear in an SNP or LSP. When the TLV appears (with a non-zero IID) in an SNP or LSP, exactly one ITID MUST be present indicating the instance-specific topology with which the PDU is associated. If no ITIDs or multiple ITIDs are present or the IID is zero, then the PDU MUST be ignored.

When the IID is non-zero and the TLV appears in an IIH, the set of ITIDs supported on the circuit over which the IIH is sent is included. There MUST be at least one ITID present.

ITID #0 is reserved for a specific use case as described later in this document. ITID #0 MUST NOT be supported in combination with any non-zero ITID. If multiple ITIDs are advertised in an IIH and one of the ITIDs is #0 then the PDU MUST be ignored.

Multiple IID-TLVs MAY appear in IIHs. If multiple IID-TLVs are present and the IID value in all IID-TLVs is not the same, then the PDU MUST be ignored.

A single IID-TLV will support advertisement of up to 126 ITIDs. If multiple IID-TLVs are present in an IIH PDU, the supported set of ITIDs is the union of all ITIDs present in all IID-TLVs.

When an LSP purge is initiated, the IID-TLV MUST be retained, but the remainder of the body of the LSP SHOULD be removed. The purge procedure is described in [RFC6233] and [RFC6232].

It is recommended that (when present) the IID-TLV(s) be the first TLVs in the PDU. This allows determination of the association of a PDU with a particular instance more quickly.

A PDU without an IID-TLV belongs to the standard instance.

2.2. Instance Membership

Each MI-RTR is configured to be participating in one or more instances of IS-IS. For each non-zero instance in which it participates, an MI-RTR marks IS-IS PDUs (IIHs, LSPs, or SNPs) generated that pertain to that instance by including the IID-TLV with the appropriate instance identifier.

2.3. Use of Authentication

When authentication is in use, the IID, if present, is first used to select the authentication configuration that is applicable. The authentication check is then performed as normal. When multiple ITIDs are supported, ITID-specific authentication MAY be used in SNPs and LSPs.

2.4. Adjacency Establishment

In order to establish adjacencies, IS-IS routers exchange IIH PDUs. Two types of adjacencies exist in IS-IS: point-to-point and broadcast. The following subsections describe the additional rules an MI-RTR MUST follow when establishing adjacencies for non-zero instances.

2.4.1. Point-to-Point Adjacencies

MI-RTRs include the IID-TLV in the point-to-point Hello PDUs associated with non-zero instances that they originate. Upon reception of an IIH, an MI-RTR inspects the received IID-TLV and if the IID matches any of the IIDs that the router supports on that circuit, normal adjacency establishment procedures are used to establish an instance-specific adjacency. Note that the absence of the IID TLV implies IID #0. For instances other than IID #0, an adjacency SHOULD NOT be established unless there is at least one ITID in common.

This extension allows an MI-RTR to establish multiple adjacencies to the same physical neighbor over a point-to-point circuit. However, as the instances are logically independent, the normal expectation of at most one neighbor on a given point-to-point circuit still applies.

2.4.2. Multi-Access Adjacencies

Multi-Access (broadcast) circuits behave differently than point-to-point in that PDUs sent by one router are visible to all routers and all routers must agree on the election of a Designated Intermediate System (DIS) independent of the set of ITIDs supported.

MI-RTRs will establish adjacencies and elect a DIS per IS-IS instance. Each MI-RTR will form adjacencies only with routers that advertise support for the instances that the local router has been configured to support on that circuit. Since an MI-RTR is not required to support all possible instances on a LAN, it's possible to elect a different DIS for different instances.

2.5. Update Process Operation

For non-zero instances, a unique Update Process exists for each supported ITID.

2.5.1. Update Process Operation on Point-to-Point Circuits

On Point-to-Point circuits -- including Point-to-Point Operation over LAN [RFC5309] -- the ITID-specific Update Process only operates on that circuit for those ITIDs that are supported by both ISs operating on the circuit.

2.5.2. Update Process Operation on Broadcast Circuits

On broadcast circuits, a single DIS is elected for each supported IID independent of the set of ITIDs advertised in LAN IIHs. This requires that the DIS generate pseudo-node LSPs for all supported ITIDs and that the Update Process for all supported ITIDs operate on the broadcast circuit. Among MI-RTRs operating on a broadcast circuit, if the set of supported ITIDs for a given non-zero IID is inconsistent, connectivity for the topology (or topologies) associated with the ITIDs not supported by some MI-RTRs can be compromised.

2.6. Interoperability Considerations

[ISO10589] requires that any TLV that is not understood is silently ignored without compromising the processing of the whole IS-IS PDU (IIH, LSP, SNP).

To a router not implementing this extension, all IS-IS PDUs received will appear to be associated with the standard instance regardless of whether an IID TLV is present in those PDUs. This can cause interoperability issues unless the mechanisms and procedures discussed below are followed.

2.6.1. Interoperability Issues on Broadcast Circuits

In order for routers to correctly interoperate with routers not implementing this extension and in order not to cause disruption, a specific and dedicated Media Access Control (MAC) address is used for

multicasting IS-IS PDUs with any non-zero IID. Each level will use a specific layer 2 multicast address. Such an address allows MI-RTRs to exchange IS-IS PDUs with non-zero IIDs without these PDUs being processed by legacy routers, and therefore no disruption is caused.

When sending SNPs, LSPs, and LAN IIHs for the standard instance (IID #0) an MI-RTR will use either the AllL1IS or AllL2IS MAC layer addresses (as defined in [ISO10589]) as the destination address. When sending SNPs, LSPs, and LAN IIHs for any non-zero IID an MI-RTR MUST use one of two new dedicated layer 2 multicast addresses (AllL1MI-ISs or AllL2MI-ISs) as the destination address. These addresses are specified in Section 6.

MI-RTRs MUST discard IS-IS PDUs received if either of the following is true:

- The destination multicast address is AllL1IS, AllL2IS, or AllIS and the PDU contains an IID-TLV.

- The destination multicast address is AllL1MI-ISs or AllL2MI-ISs and the PDU contains an IID-TLV with a zero value for the IID or has no IID-TLV.

NOTE: If the multicast addresses AllL1IS, AllL2IS, and/or AllIS are improperly used to send IS-IS PDUs for non-zero IIDs, legacy systems will interpret these PDUs as being associated with IID #0. This will cause inconsistencies in the LSDB in those routers, may incorrectly maintain adjacencies, and may lead to inconsistent DIS election.

2.6.1.1. Special Considerations When Operating in Point-to-Point Mode

When operating in point-to-point mode on a broadcast circuit [RFC5309], an MI-RTR will use AllL1IS, AllL2IS, or AllIS MAC-layer addresses when sending SNPs, LSPs, and point-to-point IIHs associated with the standard instance. When sending SNPs, LSPs, and point-to-point IIHs for a non-zero IID an MI-RTR MUST use one of the two new multicast addresses (AllL1MI-ISs or AllL2MI-IS) as the destination address. When sending point-to-point IIHs for a non-zero IID either address is permitted.

2.6.2. Interoperability Using Point-to-Point Circuits

In order for an MI-RTR to interoperate over a point-to-point circuit with a router that does NOT support this extension, the MI-RTR MUST NOT send IS-IS PDUs for instances other than IID #0 over the point-to-point circuit as these PDUs may affect the state of IID #0 in the neighbor.

The presence or absence of the IID-TLV in an IIH indicates that the neighbor does or does not support this extension, respectively. Therefore, all IIHs sent on a point-to-point circuit by an MI-RTR MUST include an IID-TLV. This includes IIHs associated with IID #0. Once it is determined that the neighbor does not support this extension, an MI-RTR MUST NOT send PDUs (including IIHs) for instances other than IID #0.

Until an IIH is received from a neighbor, an MI-RTR MAY send IIHs for a non-zero instance. However, once an IIH with no IID TLV has been received -- indicating that the neighbor is not an MI-RTR -- the MI-RTR MUST NOT send IIHs for a non-zero instance. The temporary relaxation of the restriction on sending IIHs for non-zero instances allows a non-zero instance adjacency to be established on an interface on which an MI-RTR does NOT support the standard instance.

Point-to-point adjacency setup MUST be done through the use of the three-way handshaking procedure as defined in [RFC5303] in order to prevent a non-MI capable neighbor from bringing up an adjacency prematurely based on reception of an IIH with an IID-TLV for a non-zero instance.

3. Usage Guidelines

As discussed above, MI-IS-IS extends IS-IS to support multiple instances on a given circuit. Each instance is uniquely identified by the IID and forms instance-specific adjacencies. Each instance supports one or more topologies as represented by the ITIDs. All topologies associated with a given instance share the instance-specific adjacencies. The set of topologies supported by a given IID MAY differ from circuit to circuit. Each topology has its own set of LSPs and runs a topology-specific Update Process. Flooding of topology-specific LSPs is only performed on circuits on which both the local router and the neighbor(s) support a given topology (i.e., advertise the same ITID in the set of supported ITIDs sent in the IID-TLV included in IIHs).

The following subsections provide some guidelines for usage of instances and topologies within each instance. While this represents examples based on the intent of the authors, implementors are not constrained by the examples.

3.1. One-to-One Mapping between Topologies and Instances

When the set of information to be flooded in LSPs is intended to be flooded to all MI-RTRs supporting a given IID, a single topology MAY be used. The information contained in the single LSDB MAY still contain information associated with multiple applications as the

GENINFO TLV for each application has an application-specific ID that identifies the application to which the TLV applies [RFC6823].

3.2. Many-to-One Mapping between Topologies and Instances

When the set of information to be flooded in LSPs includes subsets that are of interest to a subset of the MI-RTRs supporting a given IID, support of multiple ITIDs allows each subset to be flooded only to those MI-RTRs that are interested in that subset. In the simplest case, a one-to-one mapping between a given application and an ITID allows the information associated with that application to be flooded only to MI-RTRs that support that application -- but a many-to-one mapping between applications and a given ITID is also possible. When the set of application-specific information is large, the use of multiple ITIDs provides significantly greater efficiencies, as MI-RTRs only need to maintain the LSDB for applications of interest and that information only needs to be flooded over a topology defined by the MI-RTRs who support a given ITID.

The use of multiple ITIDs also allows the dedication of a full LSP set (256 LSPs at each level) for the use of a given (set of) applications, thereby minimizing the possibility of exceeding the carrying capacity of an LSP set. Such a possibility might arise if information for all applications were to be included in a single LSP set.

Note that the topology associated with each ITID MUST be fully connected in order for ITID-specific LSPs to be successfully flooded to all MI-RTRs that support that ITID.

When multiple ITIDs are supported by an instance ITID #0 MUST NOT be supported.

3.3. Considerations for the Number of Instances

The support of multiple topologies within the context of a single instance provides better scalability in support of multiple applications both in terms of the number of adjacencies that are required and in the flooding of topology-specific LSDB. In many cases, the use of a single non-zero instance would be sufficient and optimal. However, in cases where the set of topologies desired in support of a set of applications is largely disjoint from the set of topologies desired in support of a second set of applications, it could make sense to use multiple instances.

4. Relationship to M-ISIS

[RFC5120] defines support for multi-topology routing. In that document, 12-bit Multi-Topology Identifiers (MTIDs) are defined to identify the topologies that an IS-IS instance (a "standard instance" as defined by this document) supports. There is no relationship between the Multi-topology IDs defined in [RFC5120] and the ITIDs defined in this document.

An MI-RTR MAY use the extensions defined in this document to support multiple topologies in the context of an instance with a non-zero IID. Each MI topology is associated with a unique LSDB identified by an ITID. An ITID-specific IS-IS Update Process operates on each topology. This differs from [RFC5120] where a single LSDB and single IS-IS Update Process is used in support of all topologies. In such cases if an MI-RTR uses the extensions in support of the BFD Enabled TLV [RFC6213] , the ITID MUST be used in place of the MTID in which case all 16 bits of the identifier field are useable.

An MI-RTR MAY support [RFC5120] multi-topology within a non-zero instance when ITID #0 is supported. When ITID #0 is supported it MUST be the only ITID supported by that instance. In such cases, if an MI-RTR uses the extensions in support of the BFD Enabled TLV [RFC6213] the [RFC5120] MTID MUST be used as specified in [RFC6213].

An MI-RTR MUST NOT support [RFC5120] multi-topology within a non-zero instance when any non-zero ITID is supported. The following TLVs MUST NOT be sent in an LSP associated with a non-zero instance which supports a non-zero ITID and such an LSP MUST be ignored when received:

TLV 222 - MT IS Neighbors
TLV 235 - MT IP Reachability
TLV 237 - MT IPv6 Reachability

5. Graceful Restart Interactions

[RFC5306] defines protocol extensions in support of graceful restart of a routing instance. The extensions defined there apply to MI-RTRs with the notable addition that as there are topology-specific LSP databases all of the topology-specific LSP databases must be synchronized following restart in order for database synchronization to be complete. This involves the use of additional T2 timers. See [RFC5306] for further details.

6. IANA Considerations

Per [RFC6822], IANA has registered a new IS-IS TLV, which is reflected in the "IS-IS TLV Codepoints" registry:

Type	Description	IIH	LSP	SNP	Purge
7	Instance Identifier	y	y	y	y

Per [RFC6822], IANA has registered two EUI-48 multicast addresses from the IANA-managed EUI address space as specified in [RFC7042]. The addresses are as follows:

```
01-00-5E-90-00-02 AllL1MI-ISs
01-00-5E-90-00-03 AllL2MI-ISs
```

All references to [RFC6822] in the IS-IS TLV Codepoints registry should be replaced by references to this document.

7. Security Considerations

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

8. Acknowledgements

For the first version of this specification the authors acknowledged contributions made by Dino Farinacci and Tony Li.

For the new version of this specification the authors would like to acknowledge Paul Wells.

Most of all we would like to acknowledge Mike Shand, Abhay Roy, and Dave Ward for their contributions as co-authors of RFC 6822.

9. References

9.1. Normative References

[ISO10589] "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

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

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

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

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

[RFC5306] Shand, M. and L. Ginsberg, "Restart Signaling for IS-IS", RFC 5306, DOI 10.17487/RFC5306, October 2008, <http://www.rfc-editor.org/info/rfc5306>.

[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, <http://www.rfc-editor.org/info/rfc5310>.

[RFC6213] Hopps, C. and L. Ginsberg, "IS-IS BFD-Enabled TLV", RFC 6213, DOI 10.17487/RFC6213, April 2011, <http://www.rfc-editor.org/info/rfc6213>.

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

[RFC6233] Li, T. and L. Ginsberg, "IS-IS Registry Extension for Purges", RFC 6233, DOI 10.17487/RFC6233, May 2011, <http://www.rfc-editor.org/info/rfc6233>.

[RFC6822] Previdi, S., Ed., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, DOI 10.17487/RFC6822, December 2012, <http://www.rfc-editor.org/info/rfc6822>.

[RFC6823] Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, DOI 10.17487/RFC6823, December 2012, <http://www.rfc-editor.org/info/rfc6823>.

9.2. Informative References

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

[RFC7042] Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, DOI 10.17487/RFC7042, October 2013, <http://www.rfc-editor.org/info/rfc7042>.

Appendix A. Changes to RFC 6822

RFC 6822 prohibited the use of RFC 5120 Multi-Topology (MT) support in a non-zero instance. However, deployment experience since the writing of RFC 6822 has revealed a desire to be able to support RFC 5120 style MT using multiple non-zero instances as an alternative means of controlling leaking of information between L1 areas while also supporting incongruent topologies for different address families. The rules have therefore been relaxed to allow use of RFC 5120 MT in a non-zero instance so long as ITID #0 is the only instance topology (ITID) supported by the instance. Note that this change is not backwards compatible with implementations strictly following RFC 6822. As of this writing all known implementations are compatible with this change.

A suggestion has been added to place the IID-TLV as the first TLV in a PDU to speed recognition of the correct instance when parsing a received PDU.

Clarification that when operating in point-to-point mode on a broadcast circuit the IID-TLV is only included in Pt-Pt IIHs associated with non-zero instances has been added. This addresses Errata ID #4519.

Clarification of the appropriate MAC multicast addresses to use when sending PDUs on a broadcast interface for both standard instance and non-zero instances has been provided. This addresses Errata ID #4520.

Authors' Addresses

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

Email: ginsberg@cisco.com

Stefano Previdi
Cisco Systems
Via Del Serafico 200
Rome 0144
Italy

Email: sprevidi@cisco.com

Wim Henderickx
Nokia
BE

Email: wim.henderickx@nokia.com

IS-IS for IP Internets
Internet-Draft
Intended status: Standards Track
Expires: November 20, 2019

S. Previdi, Ed.
Huawei
L. Ginsberg, Ed.
C. Filsfils
Cisco Systems, Inc.
A. Bashandy
Arrcus
H. Gredler
RtBrick Inc.
B. Decraene
Orange

May 19, 2019

IS-IS Extensions for Segment Routing
draft-ietf-isis-segment-routing-extensions-25

Abstract

Segment Routing (SR) allows for a flexible definition of end-to-end paths within IGP topologies by encoding paths as sequences of topological sub-paths, called "segments". These segments are advertised by the link-state routing protocols (IS-IS and OSPF).

This draft describes the necessary IS-IS extensions that need to be introduced for Segment Routing operating on an MPLS data-plane.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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 November 20, 2019.

Table of Contents

1. Introduction

Segment Routing (SR) allows for a flexible definition of end-to-end paths within IGP topologies by encoding paths as sequences of topological sub-paths, called "segments". These segments are advertised by the link-state routing protocols (IS-IS and OSPF). 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's control-plane can be applied to both IPv6 and MPLS data-planes, and does not require any additional signaling (other than the regular IGP). For example, when used in MPLS networks, SR paths do not require any LDP or RSVP-TE signaling. Still, SR can interoperate in the presence of LSPs established with RSVP or LDP.

There are additional segment types, e.g., Binding SID defined in [RFC8402]. This document also defines an advertisement for one type of Binding SID: the Mirror Context segment.

This draft describes the necessary IS-IS extensions that need to be introduced for Segment Routing operating on an MPLS data-plane.

The Segment Routing architecture is described in [RFC8402].

Segment Routing use cases are described in [RFC7855].

2. Segment Routing Identifiers

The Segment Routing architecture [RFC8402] defines different types of Segment Identifiers (SID). This document defines the IS-IS encodings for the IGP-Prefix Segment, the IGP-Adjacency Segment, the IGP-LAN-Adjacency Segment and the Binding Segment.

2.1. Prefix Segment Identifier (Prefix-SID Sub-TLV)

A new IS-IS sub-TLV is defined: the Prefix Segment Identifier sub-TLV (Prefix-SID sub-TLV).

The Prefix-SID sub-TLV carries the Segment Routing IGP-Prefix-SID as defined in [RFC8402]. The 'Prefix SID' MUST be unique within a given IGP domain (when the L-flag is not set).

A Prefix-SID sub-TLV is associated to a prefix advertised by a node and MAY be present in any of the following TLVs:

TLV-135 (Extended IPv4 reachability) defined in [RFC5305].

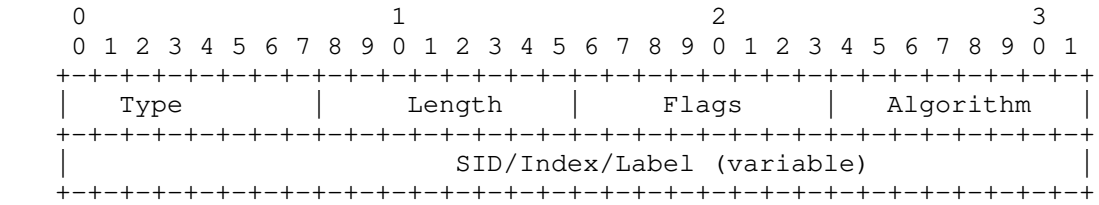
TLV-235 (Multitopology IPv4 Reachability) defined in [RFC5120].

TLV-236 (IPv6 IP Reachability) defined in [RFC5308].

TLV-237 (Multitopology IPv6 IP Reachability) defined in [RFC5120].

Binding-TLV and Multi-Topology Binding-TLV defined in Section 2.4 and Section 2.5 respectively.

The Prefix-SID sub-TLV has the following format:

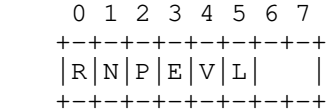


where:

Type: 3

Length: 5 or 6 depending on the size of the SID (described below)

Flags: 1 octet field of following flags:



where:

R-Flag: Re-advertisement flag. If set, then the prefix to which this Prefix-SID is attached, has been propagated by the router either from another level (i.e., from level-1 to level-2 or the opposite) or from redistribution (e.g.: from another protocol).

N-Flag: Node-SID flag. If set, then the Prefix-SID refers to the router identified by the prefix. Typically, the N-Flag is set on Prefix-SIDs attached to a router loopback address. The N-Flag is set when the Prefix-SID is a Node-SID as described in [RFC8402].

P-Flag: no-PHP flag. If set, then the penultimate hop MUST NOT pop the Prefix-SID before delivering the packet to the node that advertised the Prefix-SID.

E-Flag: Explicit-Null Flag. If set, any upstream neighbor of the Prefix-SID originator MUST replace the Prefix-SID with a Prefix-SID having an Explicit-NULL value (0 for IPv4 and 2 for IPv6) before forwarding the packet.

V-Flag: Value flag. If set, then the Prefix-SID carries a value (instead of an index). By default the flag is UNSET.

L-Flag: Local Flag. If set, then the value/index carried by the Prefix-SID has local significance. By default the flag is UNSET.

Other bits: MUST be zero when originated and ignored when received.

Algorithm: the router may use various algorithms when calculating reachability to other nodes or to prefixes attached to these nodes. Algorithm identifiers are defined in Section 3.2. Examples of these algorithms are metric based Shortest Path First (SPF), various sorts of Constrained SPF, etc. The algorithm field of the Prefix-SID contains the identifier of the algorithm the router uses to compute the reachability of the prefix to which the Prefix-SID is associated.

At origination, the Prefix-SID algorithm field MUST be set to 0 or to any value advertised in the SR-Algorithm sub-TLV (Section 3.2).

A router receiving a Prefix-SID from a remote node and with an algorithm value that such remote node has not advertised in the SR-Algorithm sub-TLV (Section 3.2) MUST ignore the Prefix-SID sub-TLV.

SID/Index/Label as defined in Section 2.1.1.1.

When the Prefix SID is an index (the V-flag is not set) the value is used to determine the actual label value inside the set of all advertised label ranges of a given router. This allows a receiving router to construct forwarding state to a particular destination router.

In many use-cases a 'stable transport' address is overloaded as an identifier of a given node. Because Prefixes may be re-advertised into other levels there may be some ambiguity (e.g. Originating router vs. L1L2 router) for which node a particular IP prefix serves as identifier. The Prefix-SID sub-TLV contains the necessary flags to disambiguate Prefix to node mappings. Furthermore if a given node has several 'stable transport' addresses there are flags to differentiate those among other Prefixes advertised from a given node.

2.1.1. Flags

2.1.1.1. V and L Flags

The V-flag indicates whether the SID/Index/Label field is a value or an index.

The L-Flag indicates whether the value/index in the SID/Index/Label field has local or global significance.

The following settings for V and L flags are valid:

V-flag is set to 0 and L-flag is set to 0: The SID/Index/Label field is a 4 octet index defining the offset in the SID/Label space advertised by this router using the encodings defined in Section 3.1.

V-flag is set to 1 and L-flag is set to 1: The SID/Index/Label field is a 3 octet local label where the 20 rightmost bits are used for encoding the label value.

All other combinations of V-flag and L-flag are invalid and any SID advertisement received with an invalid setting for V and L flags MUST be ignored.

2.1.1.2. R and N Flags

The R-Flag MUST be set for prefixes that are not local to the router and either:

advertised because of propagation (Level-1 into Level-2);

advertised because of leaking (Level-2 into Level-1);

advertised because of redistribution (e.g.: from another protocol).

In the case where a Level-1-2 router has local interface addresses configured in one level, it may also propagate these addresses into the other level. In such case, the Level-1-2 router MUST NOT set the R bit.

The N-Flag is used in order to define a Node-SID. A router MAY set the N-Flag only if all of the following conditions are met:

The prefix to which the Prefix-SID is attached is local to the router (i.e., the prefix is configured on one of the local interfaces, e.g., a 'stable transport' loopback).

The prefix to which the Prefix-SID is attached has a Prefix length of either /32 (IPv4) or /128 (IPv6).

The router MUST ignore the N-Flag on a received Prefix-SID if the prefix has a Prefix length different than /32 (IPv4) or /128 (IPv6).

The Prefix Attributes Flags sub-TLV [RFC7794] also defines the N and R flags and with the same semantics of the equivalent flags defined in this document. Whenever the Prefix Attributes Flags sub-TLV is present for a given prefix the values of the N and R flags advertised in that sub-TLV MUST be used and the values in a corresponding Prefix SID sub-TLV (if present) MUST be ignored.

2.1.1.3. E and P Flags

The following behavior is associated with the settings of the E and P flags:

- If the P-flag is not set then any upstream neighbor of the Prefix-SID originator MUST pop the Prefix-SID. This is equivalent to the penultimate hop popping mechanism used in the MPLS dataplane which improves performance of the ultimate hop. MPLS EXP bits of the Prefix-SID are not preserved to the ultimate hop (the Prefix-SID being removed). If the P-flag is unset the received E-flag is ignored.

- If the P-flag is set then:

  * If the E-flag is not set then any upstream neighbor of the Prefix-SID originator MUST keep the Prefix-SID on top of the stack. This is useful when, e.g., the originator of the

Prefix-SID must stitch the incoming packet into a continuing MPLS LSP to the final destination. This could occur at an inter-area border router (prefix propagation from one area to another) or at an inter-domain border router (prefix propagation from one domain to another).

* If the E-flag is set then any upstream neighbor of the Prefix-SID originator MUST replace the PrefixSID with a Prefix-SID having an Explicit-NULL value. This is useful, e.g., when the originator of the Prefix-SID is the final destination for the related prefix and the originator wishes to receive the packet with the original EXP bits.

When propagating (either from Level-1 to Level-2 or vice versa) a reachability advertisement originated by another IS-IS speaker, the router MUST set the P-flag and MUST clear the E-flag of the related Prefix-SIDs.

2.1.2. Prefix-SID Propagation

The Prefix-SID sub-TLV MUST be included when the associated Prefix Reachability TLV is propagated across level boundaries.

The level-1-2 router that propagates the Prefix-SID sub-TLV between levels maintains the content (flags and SID) except as noted in Section 2.1.1.2 and Section 2.1.1.3.

2.2. Adjacency Segment Identifier

A new IS-IS sub-TLV is defined: the Adjacency Segment Identifier sub-TLV (Adj-SID sub-TLV).

The Adj-SID sub-TLV is an optional sub-TLV carrying the Segment Routing IGP-Adjacency-SID as defined in [RFC8402] with flags and fields that may be used, in future extensions of Segment Routing, for carrying other types of SIDs.

IS-IS adjacencies are advertised using one of the IS-Neighbor TLVs below:

TLV-22 (Extended IS reachability)[RFC5305]

TLV-222 (Multitopology IS)[RFC5120]

TLV-23 (IS Neighbor Attribute)[RFC5311]

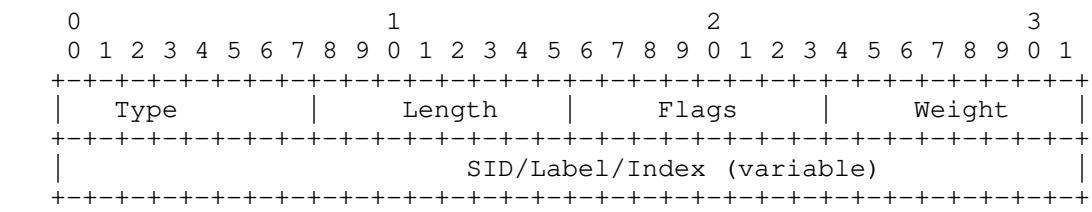
TLV-223 (Multitopology IS Neighbor Attribute)[RFC5311]

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

Multiple Adj-SID sub-TLVs MAY be associated with a single IS-neighbor.

2.2.1. Adjacency Segment Identifier (Adj-SID) Sub-TLV

The following format is defined for the Adj-SID sub-TLV:

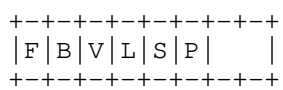


where:

Type: 31

Length: 5 or 6 depending on size of the SID

Flags: 1 octet field of following flags:



where:

F-Flag: Address-Family flag. If unset, then the Adj-SID is used when forwarding IPv4 encapsulated traffic to the neighbor. If set then the Adj-SID is used when forwarding IPv6 encapsulated traffic to the neighbor.

B-Flag: Backup flag. If set, the Adj-SID is eligible for protection (e.g.: using IPFRR or MPLS-FRR) as described in [RFC8402].

V-Flag: Value flag. If set, then the Adj-SID carries a value. By default the flag is SET.

L-Flag: Local Flag. If set, then the value/index carried by the Adj-SID has local significance. By default the flag is SET.

S-Flag. Set flag. When set, the S-Flag indicates that the Adj-SID refers to a set of adjacencies (and therefore MAY be assigned to other adjacencies as well).

P-Flag. Persistent flag. When set, the P-Flag indicates that the Adj-SID is persistently allocated, i.e., the Adj-SID value remains consistent across router restart and/or interface flap.

Other bits: MUST be zero when originated and ignored when received.

Weight: 1 octet. The value represents the weight of the Adj-SID for the purpose of load balancing. The use of the weight is defined in [RFC8402].

SID/Index/Label as defined in Section 2.1.1.1.

An SR capable router MAY allocate an Adj-SID for each of its adjacencies

An SR capable router MAY allocate more than one Adj-SID to an adjacency.

An SR capable router MAY allocate the same Adj-SID to different adjacencies.

When the P-flag is not set, the Adj-SID MAY be persistent. When the P-flag is set, the Adj-SID MUST be persistent.

Examples of use of the Adj-SID sub-TLV are described in [RFC8402].

The F-flag is used in order for the router to advertise the outgoing encapsulation of the adjacency the Adj-SID is attached to.

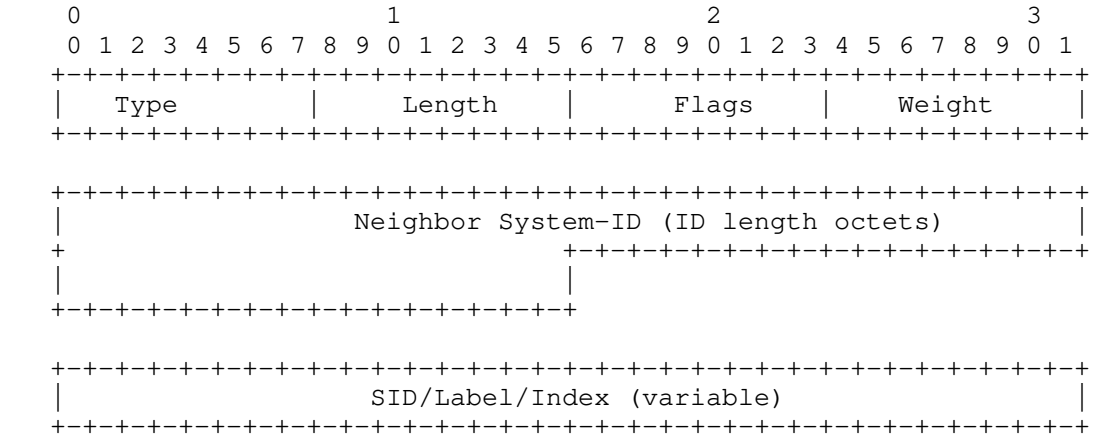
2.2.2. Adjacency Segment Identifiers in LANs

In LAN subnetworks, the Designated Intermediate System (DIS) is elected and originates the Pseudonode-LSP (PN-LSP) including all neighbors of the DIS.

When Segment Routing is used, each router in the LAN MAY advertise the Adj-SID of each of its neighbors. Since, on LANs, each router only advertises one adjacency to the DIS (and doesn't advertise any other adjacency), each router advertises the set of Adj-SIDs (for each of its neighbors) inside a newly defined sub-TLV part of the TLV advertising the adjacency to the DIS (e.g.: TLV-22).

The following new sub-TLV is defined: LAN-Adj-SID containing the set of Adj-SIDs the router assigned to each of its LAN neighbors.

The format of the LAN-Adj-SID sub-TLV is as follows:

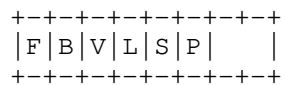


where:

Type: 32

Length: variable.

Flags: 1 octet field of following flags:



where F, B, V, L, S and P flags are defined in Section 2.2.1. Other bits: MUST be zero when originated and ignored when received.

Weight: 1 octet. The value represents the weight of the Adj-SID for the purpose of load balancing. The use of the weight is defined in [RFC8402].

Neighbor System-ID: IS-IS System-ID of length "ID Length" as defined in [ISO10589].

SID/Index/Label as defined in Section 2.1.1.1.

Multiple LAN-Adj-SID sub-TLVs MAY be encoded.

Note that this sub-TLV MUST NOT appear in TLV 141.

In case one TLV-22/23/222/223 (reporting the adjacency to the DIS) can't contain the whole set of LAN-Adj-SID sub-TLVs, multiple advertisements of the adjacency to the DIS MUST be used and all advertisements MUST have the same metric.

Each router within the level, by receiving the DIS PN LSP as well as the non-PN LSP of each router in the LAN, is capable of reconstructing the LAN topology as well as the set of Adj-SIDs each router uses for each of its neighbors.

2.3. SID/Label Sub-TLV

The SID/Label sub-TLV may be present in the following TLVs/sub-TLVs defined in this document:

SR-Capabilities Sub-TLV (Section 3.1)

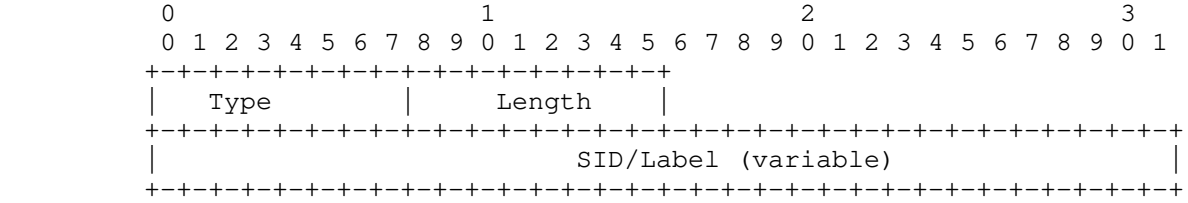
SR Local Block Sub-TLV (Section 3.3)

SID/Label Binding TLV (Section 2.4)

Multi-Topology SID/Label Binding TLV (Section 2.5)

Note that the code point used in all of the above cases is the SID/Label Sub-TLV code point specified in the new "sub-TLVs for TLV 149 and 150" registry created by this document.

The SID/Label sub-TLV contains a SID or a MPLS Label. The SID/Label sub-TLV has the following format:



where:

Type: 1

Length: 3 or 4

SID/Label: if length is set to 3 then the 20 rightmost bits represent a MPLS label. If length is set to 4 then the value is a 32 bit index

2.4. SID/Label Binding TLV

The SID/Label Binding TLV MAY be originated by any router in an IS-IS domain. There are multiple uses of the SID/Label Binding TLV.

The SID/Label Binding TLV may be used to advertise prefixes to SID/Label mappings. This functionality is called the Segment Routing Mapping Server (SRMS). The behavior of the SRMS is defined in [I-D.ietf-spring-segment-routing-ldp-interop].

The SID/Label Binding TLV may also be used to advertise a Mirror SID to advertise the ability to process traffic originally destined to another IGP node. This behavior is defined in [RFC8402].

The SID/Label Binding TLV has the following format:

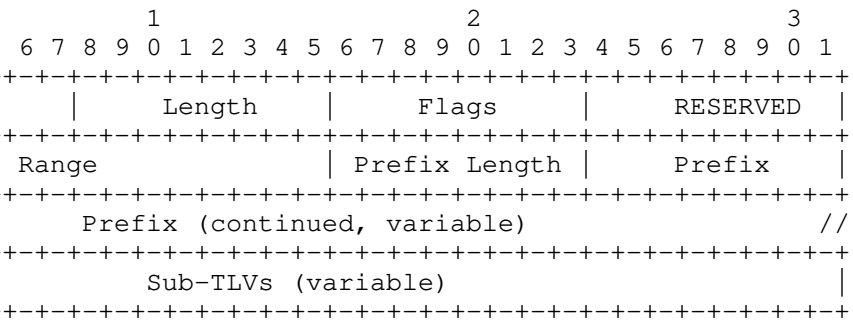


Figure 1: SID/Label Binding TLV format

- Type: 149
- Length: variable.
- 1 octet of flags
- 1 octet of RESERVED (SHOULD be transmitted as 0 and MUST be ignored on receipt)
- 2 octets of Range
- 1 octet of Prefix Length
- 0-16 octets of Prefix

o sub-TLVs, where each sub-TLV consists of a sequence of:

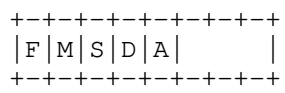
* 1 octet of sub-TLV type

* 1 octet of length of the value field of the sub-TLV

* 0-243 octets of value

2.4.1. Flags

Flags: 1 octet field of following flags:



where:

F-Flag: Address Family flag. If unset, then the Prefix carries an IPv4 Prefix. If set then the Prefix carries an IPv6 Prefix.

M-Flag: Mirror Context flag. Set if the advertised SID corresponds to a mirrored context. The use of a mirrored context is described in [RFC8402].

S-Flag: If set, the SID/Label Binding TLV SHOULD be flooded across the entire routing domain. If the S flag is not set, the SID/ Label Binding TLV MUST NOT be leaked between levels. This bit MUST NOT be altered during the TLV leaking.

D-Flag: when the SID/Label Binding TLV is leaked from level-2 to level-1, the D-Flag MUST be set. Otherwise, this flag MUST be clear. SID/Label Binding TLVs with the D-Flag set MUST NOT be leaked from level-1 to level-2. This is to prevent TLV looping across levels.

A-Flag: Attached flag. The originator of the SID/Label Binding TLV MAY set the A bit in order to signal that the prefixes and SIDs advertised in the SID/Label Binding TLV are directly connected to their originators. The mechanisms through which the originator of the SID/Label Binding TLV can figure out if a prefix is attached or not are outside the scope of this document (e.g.: through explicit configuration). If the Binding TLV is leaked to other areas/levels the A-flag MUST be cleared.

An implementation may decide not to honor the S-flag in order not to leak Binding TLV's between levels (for policy reasons).

Other bits: MUST be zero when originated and ignored when received.

2.4.2. Range

The 'Range' field provides the ability to specify a range of addresses and their associated Prefix SIDs. This advertisement supports the SRMS functionality. It is essentially a compression scheme to distribute a continuous Prefix and their continuous, corresponding SID/Label Block. If a single SID is advertised then the range field MUST be set to one. For range advertisements > 1, the range field MUST be set to the number of addresses that need to be mapped into a Prefix-SID. In either case the prefix is the first address to which a SID is to be assigned.

2.4.3. Prefix Length, Prefix

The 'Prefix' represents the Forwarding equivalence class at the tail-end of the advertised path. The 'Prefix' does not need to correspond to a routable prefix of the originating node.

The 'Prefix Length' field contains the length of the prefix in bits. Only the most significant octets of the Prefix are encoded (i.e., 1 octet for prefix length 1 up to 8, 2 octets for prefix length 9 to 16, 3 octets for prefix length 17 up to 24 and 4 octets for prefix length 25 up to 32, ..., 16 octets for prefix length 113 up to 128).

2.4.4. Mapping Server Prefix-SID

The Prefix-SID sub-TLV is defined in Section 2.1 and contains the SID/index/label value associated with the prefix and range. The Prefix-SID Sub-TLV MUST be present in the SID/Label Binding TLV when the M-flag is clear. The Prefix-SID Sub-TLV MUST NOT be present when the M-flag is set.

2.4.4.1. Prefix-SID Flags

The Prefix-SID flags are defined in Section 2.1. The Mapping Server MAY advertise a mapping with the N flag set when the prefix being mapped is known in the link-state topology with a mask length of 32 (IPv4) or 128 (IPv6) and when the prefix represents a node. The mechanisms through which the operator defines that a prefix represents a node are outside the scope of this document (typically it will be through configuration).

The other flags defined in Section 2.1 are not used by the Mapping Server and MUST be ignored at reception.

2.4.4.2. PHP Behavior when using Mapping Server Advertisements

As the mapping server does not specify the originator of a prefix advertisement it is not possible to determine PHP behavior solely based on the Mapping Server Advertisement. However, if additional information is available PHP behavior may safely be done. The required information consists of:

- A prefix reachability advertisement for the prefix has been received which includes the Prefix Attribute Flags sub-TLV [RFC7794].
- X and R flags are both set to 0 in the Prefix Attribute Flags sub-TLV.

In the absence of an Prefix Attribute Flags sub-TLV [RFC7794] the A flag in the binding TLV indicates that the originator of a prefix reachability advertisement is directly connected to the prefix and thus PHP MUST be done by the neighbors of the router originating the prefix reachability advertisement. Note that A-flag is only valid in the original area in which the Binding TLV is advertised.

2.4.4.3. Prefix-SID Algorithm

The algorithm field contains the identifier of the algorithm associated with the SIDs for the prefix(es) in the range. Use of the algorithm field is described in Section 2.1.

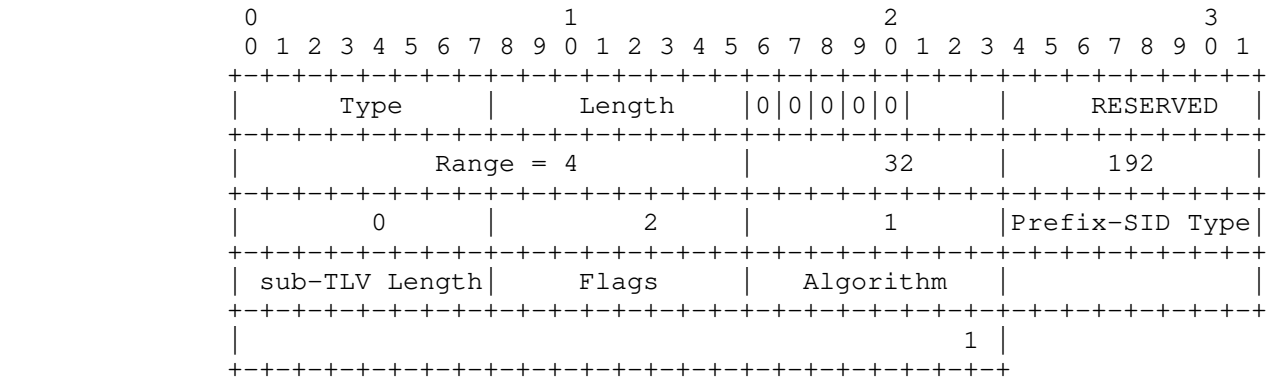
2.4.5. SID/Label Sub-TLV

The SID/Label sub-TLV (Type: 1) contains the SID/Label value as defined in Section 2.3. It MUST be present in the SID/Label Binding TLV when the M-flag is set in the Flags field of the parent TLV.

2.4.6. Example Encodings

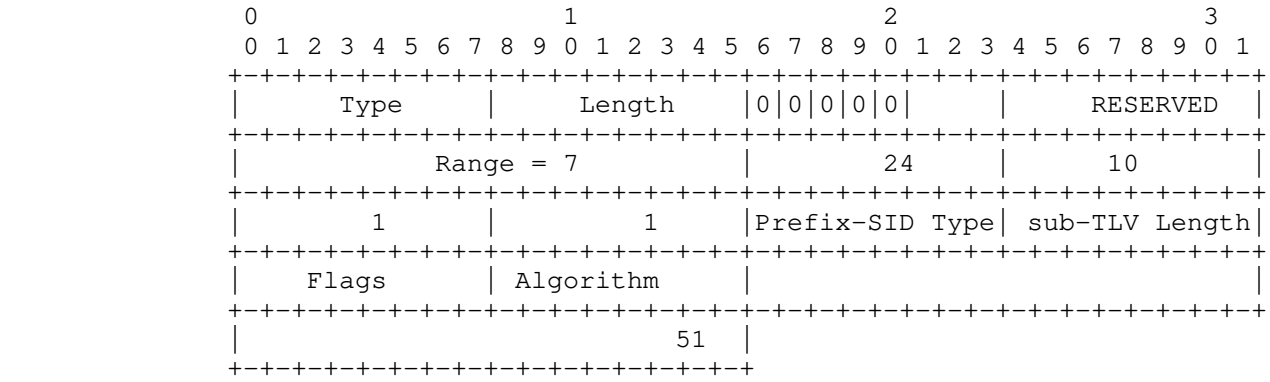
Example 1: if the following IPv4 router addresses (loopback addresses) need to be mapped into the corresponding Prefix SID indexes.

Router-A: 192.0.2.1/32, Prefix-SID: Index 1
Router-B: 192.0.2.2/32, Prefix-SID: Index 2
Router-C: 192.0.2.3/32, Prefix-SID: Index 3
Router-D: 192.0.2.4/32, Prefix-SID: Index 4



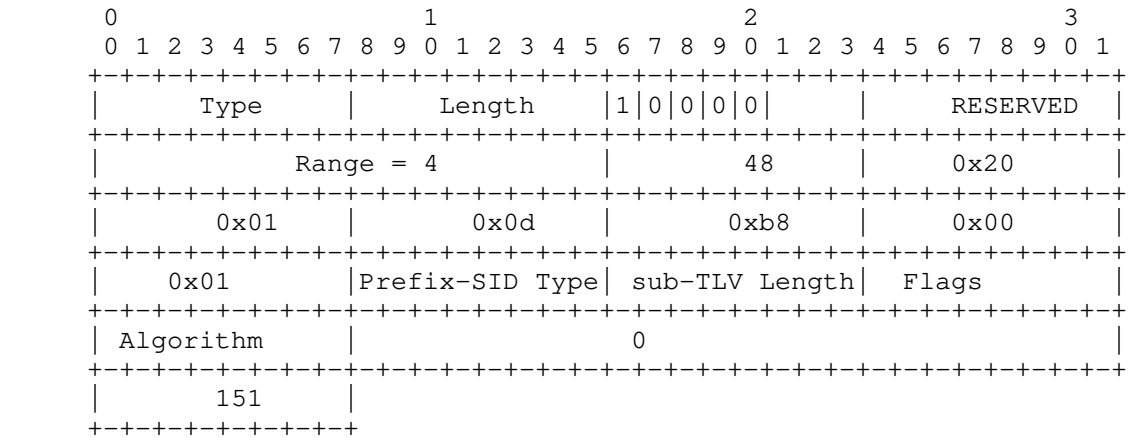
Example-2: If the following IPv4 prefixes need to be mapped into the corresponding Prefix-SID indexes:

10.1.1/24, Prefix-SID: Index 51
10.1.2/24, Prefix-SID: Index 52
10.1.3/24, Prefix-SID: Index 53
10.1.4/24, Prefix-SID: Index 54
10.1.5/24, Prefix-SID: Index 55
10.1.6/24, Prefix-SID: Index 56
10.1.7/24, Prefix-SID: Index 57



Example-3: If the following IPv6 prefixes need to be mapped into the corresponding Prefix-SID indexes:

2001:db8:1/48, Prefix-SID: Index 151
2001:db8:2/48, Prefix-SID: Index 152
2001:db8:3/48, Prefix-SID: Index 153
2001:db8:4/48, Prefix-SID: Index 154



It is not expected that a network operator will be able to keep fully continuous Prefix / SID/Index mappings. In order to support noncontinuous mapping ranges an implementation MAY generate several instances of Binding TLVs.

For example if a router wants to advertise the following ranges:

Range 16: { 192.0.2.1-15, Index 1-15 }

Range 6: { 192.0.2.22-27, Index 22-27 }

Range 41: { 192.0.2.44-84, Index 80-120 }

A router would need to advertise three instances of the Binding TLV.

2.5. Multi-Topology SID/Label Binding TLV

The Multi-Topology SID/Label Binding TLV allows the support of M-ISIS as defined in [RFC5120]. The Multi-Topology SID/Label Binding TLV has the same format as the SID/Label Binding TLV defined in Section 2.4 with the difference consisting of a Multitopology Identifier (MTID) as defined here below:

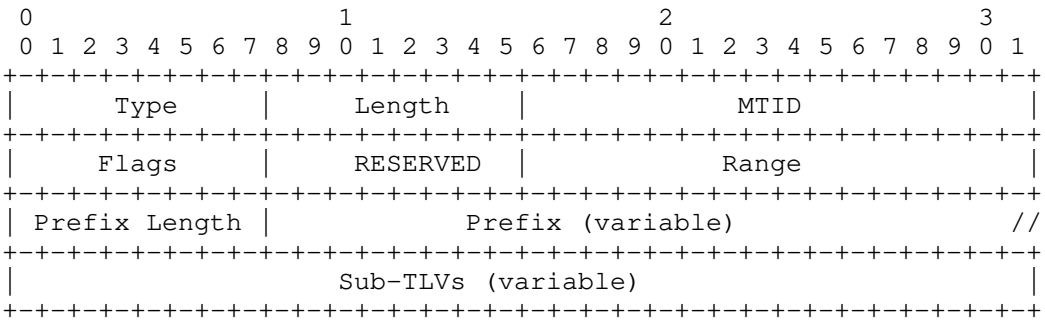


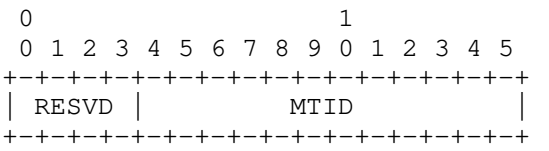
Figure 2: Multi-Topology SID/Label Binding TLV format

where:

Type: 150

Length: variable

MTID is the multitopology identifier defined as:



RESVD: reserved bits. MUST be reset on transmission and ignored on receive.

MTID: a 12-bit field containing the non-zero ID of the topology being announced. The TLV MUST be ignored if the ID is zero. This is to ensure the consistent view of the standard unicast topology.

The other fields and Sub-TLVs are defined in Section 2.4.

3. Router Capabilities

This section defines sub-TLVs which are inserted into the IS-IS Router Capability TLV-242 that is defined in [RFC7981].

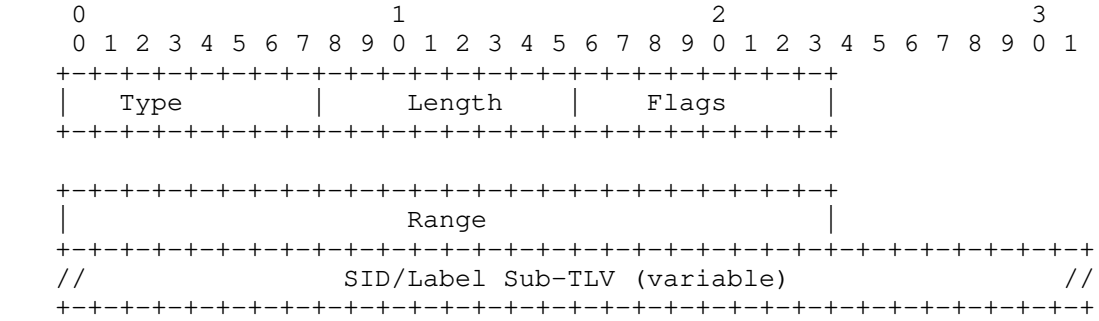
3.1. SR-Capabilities Sub-TLV

Segment Routing requires each router to advertise its SR data-plane capability and the range of MPLS label values it uses for Segment Routing in the case where global SIDs are allocated (i.e., global

indexes). Data-plane capabilities and label ranges are advertised using the newly defined SR-Capabilities sub-TLV.

The Router Capability TLV specifies flags that control its advertisement. The SR Capabilities sub-TLV MUST be propagated throughout the level and MUST NOT be advertised across level boundaries. Therefore Router Capability TLV distribution flags are set accordingly, i.e., the S flag in the Router Capability TLV [RFC7981] MUST be unset.

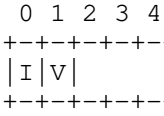
The SR Capabilities sub-TLV has following format:



Type: 2

Length: variable.

Flags: 1 octet of flags. The following are defined:



where:

I-Flag: MPLS IPv4 flag. If set, then the router is capable of processing SR MPLS encapsulated IPv4 packets on all interfaces.

V-Flag: MPLS IPv6 flag. If set, then the router is capable of processing SR MPLS encapsulated IPv6 packets on all interfaces.

One or more SRGB Descriptor entries, each of which have the following format:

Range: 3 octets.

SID/Label sub-TLV (as defined in Section 2.3).

SID/Label sub-TLV contains the first value of the SRGB while the range contains the number of SRGB elements. The range value MUST be higher than 0.

The SR-Capabilities sub-TLV MAY be advertised in an LSP of any number but a router MUST NOT advertise more than one SR-Capabilities sub-TLV. A router receiving multiple SR-Capabilities sub-TLVs from the same originator SHOULD select the first advertisement in the lowest numbered LSP.

When multiple SRGB Descriptors are advertised the entries define an ordered set of ranges on which a SID index is to be applied. For this reason changing the order in which the descriptors are advertised will have a disruptive effect on forwarding.

When a router adds a new SRGB Descriptor to an existing SR-Capabilities sub-TLV the new Descriptor SHOULD add the newly configured block at the end of the sub-TLV and SHOULD NOT change the order of previously advertised blocks. Changing the order of the advertised descriptors will create label churn in the FIB and blackhole / misdirect some traffic during the IGP convergence. In particular, if a range which is not the last is extended it's preferable to add a new range rather than extending the previously advertised range.

The originating router MUST ensure the order is unchanged after a graceful restart (using checkpointing, non-volatile storage or any other mechanism).

The originating router MUST NOT advertise overlapping ranges.

When a router receives multiple overlapping ranges, it MUST conform to the procedures defined in [I-D.ietf-spring-segment-routing-mpls].

Here follows an example of advertisement of multiple ranges:

The originating router advertises following ranges:

```
SR-Cap: range: 100, SID value: 100
SR-Cap: range: 100, SID value: 1000
SR-Cap: range: 100, SID value: 500
```

The receiving routers concatenate the ranges in the received order and build the SRGB as follows:

```
SRGB = [100, 199]
       [1000, 1099]
       [500, 599]
```

The indexes span multiple ranges:

```
index=0   means label 100
...
index 99  means label 199
index 100 means label 1000
index 199 means label 1099
...
index 200 means label 500
...
```

3.2. SR-Algorithm Sub-TLV

The router may use various algorithms when calculating reachability to other nodes or to prefixes attached to these nodes. Examples of these algorithms are metric based Shortest Path First (SPF), various sorts of Constrained SPF, etc. The SR-Algorithm sub-TLV allows the router to advertise the algorithms that the router is currently using. Algorithm values are defined in the "IGP Algorithm Type" registry defined in [I-D.ietf-ospf-segment-routing-extensions]. The following values have been defined:

0: Shortest Path First (SPF) algorithm based on link metric. This is the well-known shortest path algorithm as computed by the IS-IS Decision process. Consistent with the deployed practice for link-state protocols, algorithm 0 permits any node to overwrite the SPF path with a different path based on local policy.

1: Strict Shortest Path First (SPF) algorithm based on link metric. The algorithm is identical to algorithm 0 but algorithm 1 requires that all nodes along the path will honor the SPF routing decision. Local policy MUST NOT alter the forwarding decision computed by algorithm 1 at the node claiming to support algorithm 1.

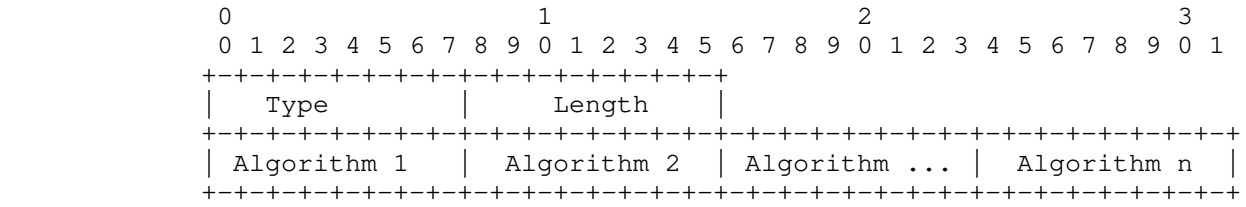
The Router Capability TLV specifies flags that control its advertisement. The SR-Algorithm MUST be propagated throughout the level and MUST NOT be advertised across level boundaries. Therefore Router Capability TLV distribution flags are set accordingly, i.e., the S flag MUST be unset.

The SR-Algorithm sub-TLV is optional. It MUST NOT be advertsied more than once at a given level. A router receiving multiple SR-Algorithm sub-TLVs from the same originator SHOULD select the first advertisement in the lowest numbered LSP.

When the originating router does not advertise the SR-Algorithm sub-TLV, this implies that the only algorithm supported by routers supporting the extensions defined in this document is Algorithm 0.

When the originating router does advertise the SR-Algorithm sub-TLV, then algorithm 0 MUST be present while non-zero algorithms MAY be present.

The SR-Algorithm sub-TLV has the following format:



where:

Type: 19

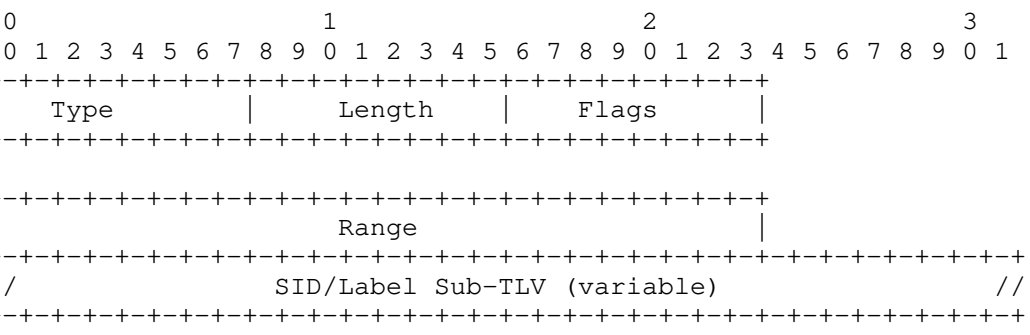
Length: variable.

Algorithm: 1 octet of algorithm

3.3. SR Local Block Sub-TLV

The SR Local Block (SRLB) Sub-TLV contains the range of labels the node has reserved for local SIDs. Local SIDs are used, e.g., for Adjacency-SIDs, and may also be allocated by components other than the IS-IS protocol. As an example, an application or a controller may instruct the router to allocate a specific local SID. Therefore, in order for such applications or controllers to know what are the local SIDs available in the router, it is required that the router advertises its SRLB.

The SRLB Sub-TLV is used for this purpose and has following format:



Type: 22

Length: variable.

Flags: 1 octet of flags. None are defined at this stage.

One or more SRLB Descriptor entries, each of which have the following format:

Range: 3 octets.

SID/Label sub-TLV (as defined in Section 2.3).

SID/Label sub-TLV contains the first value of the SRLB while the range contains the number of SRLB elements. The range value MUST be higher than 0.

The SRLB sub-TLV MAY be advertised in an LSP of any number but a router MUST NOT advertise more than one SRLB sub-TLV. A router receiving multiple SRLB sub-TLVs, from the same originator, SHOULD select the first advertisement in the lowest numbered LSP.

The originating router MUST NOT advertise overlapping ranges.

When a router receives multiple overlapping ranges, it MUST conform to the procedures defined in [I-D.ietf-spring-segment-routing-mpls].

It is important to note that each time a SID from the SRLB is allocated, it should also be reported to all components (e.g.: controller or applications) in order for these components to have an up-to-date view of the current SRLB allocation and in order to avoid collision between allocation instructions.

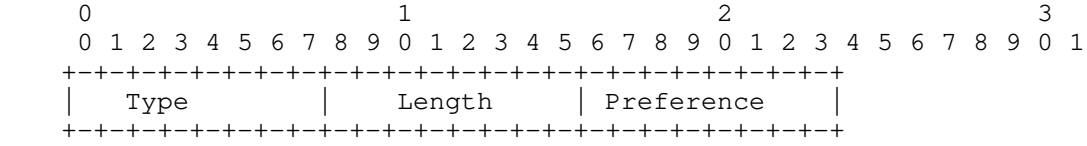
Within the context of IS-IS, the reporting of local SIDs is done through IS-IS Sub-TLVs such as the Adjacency-SID. However, the reporting of allocated local SIDs may also be done through other means and protocols which are outside the scope of this document.

A router advertising the SRLB sub-TLV may also have other label ranges, outside the SRLB, for its local allocation purposes which are NOT advertised in the SRLB. For example, it is possible that an Adjacency-SID is allocated using a local label not part of the SRLB.

3.4. SRMS Preference Sub-TLV

The Segment Routing Mapping Server (SRMS) Preference sub-TLV is used in order to associate a preference with SRMS advertisements from a particular source.

The SRMS Preference sub-TLV has following format:



Type: 24

Length: 1.

Preference: 1 octet. Unsigned 8 bit SRMS preference.

The SRMS Preference sub-TLV MAY be advertised in an LSP of any number but a router MUST NOT advertise more than one SRMS Preference sub-TLV. A router receiving multiple SRMS Preference sub-TLVs, from the same originator, SHOULD select the first advertisement in the lowest numbered LSP.

The use of the SRMS Preference during the SID selection process is described in [I-D.ietf-spring-segment-routing-ldp-interop]

4. IANA Considerations

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

4.1. Sub TLVs for Type 22,23,25,141,222, and 223

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

Type	Description	22	23	25	141	222	223
31	Adjacency Segment Identifier	y	y	n	y	y	y
32	LAN Adjacency Segment Identifier	y	y	n	y	y	y

4.2. Sub TLVs for Type 135,235,236 and 237

This document makes the following registrations in the "sub-TLVs for TLV 135,235,236 and 237" registry.

Type	Description	135	235	236	237
3	Prefix Segment Identifier	y	y	y	y

4.3. Sub TLVs for Type 242

This document makes the following registrations in the "sub-TLVs for TLV 242" registry.

Type	Description
2	Segment Routing Capability
19	Segment Routing Algorithm
22	Segment Routing Local Block (SRLB)
24	Segment Routing Mapping Server Preference (SRMS Preference)

4.4. New TLV Codepoint and Sub-TLV registry

This document registers the following TLV:

Value	Name	IIH	LSP	SNP	Purge
149	Segment Identifier/Label Binding	n	y	n	n
150	Multi-Topology Segment Identifier /Label Binding	n	y	n	n

This document creates the following sub-TLV Registry:

Name: sub-TLVs for TLVs 149 and 150
Registration Procedure: Expert Review

Type	Description
0	Reserved
1	SID/Label
2	Unassigned
3	Prefix SID
4-255	Unassigned

5. Security Considerations

With the use of the extensions defined in this document, IS-IS carries information which will be used to program the MPLS data plane [RFC3031]. In general, the same types of attacks that can be carried out on the IP/IPv6 control plane can be carried out on the MPLS control plane resulting in traffic being misrouted in the respective data planes. However, the latter may be more difficult to detect and isolate.

Existing security extensions as described in [RFC5304] and [RFC5310] apply to these segment routing extensions.

6. Acknowledgements

We would like to thank Dave Ward, Dan Frost, Stewart Bryant, Pierre Francois and Jesper Skrivers for their contribution to the content of this document.

7. Contributors

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

Stephane Litkowski
Orange
FR

Email: stephane.litkowski@orange.com

Jeff Tantsura
Apstra, Inc.

Email: jefftant@gmail.com

Peter Psenak

Cisco Systems Inc.
US

Email: ppsenak@cisco.com

Martin Horneffer
Deutsche Telekom
DE

Email: Martin.Horneffer@telekom.de

Wim Henderickx
Nokia
BE

Email: wim.henderickx@nokia.com

Edward Crabbe
Oracle
US

Email: edward.crabbe@oracle.com

Rob Shakir
Google
UK

Email: robjs@google.com

Igor Milojevic
Individual
RS

Email: milojevicigor@gmail.com

Saku Ytti
TDC
FI

Email: saku@ytti.fi

Steven Luong
Cisco Systems Inc.

US

Email: sluong@cisco.com

8. References

8.1. Normative References

[I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-27 (work in progress), December 2018.

[I-D.ietf-spring-segment-routing-ldp-interop]
Bashandy, A., Filsfils, C., Previdi, S., Decraene, B., and S. Litkowski, "Segment Routing interworking with LDP", draft-ietf-spring-segment-routing-ldp-interop-15 (work in progress), September 2018.

[I-D.ietf-spring-segment-routing-mpls]
Bashandy, A., Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-22 (work in progress), May 2019.

[ISO10589]
International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, Nov 2002.

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

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

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

[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>.

[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>.

[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>.

[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>.

8.2. Informative 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>.

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

[RFC5311] McPherson, D., Ed., Ginsberg, L., Previdi, S., and M. Shand, "Simplified Extension of Link State PDU (LSP) Space for IS-IS", RFC 5311, DOI 10.17487/RFC5311, February 2009, <https://www.rfc-editor.org/info/rfc5311>.

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

[RFC7855] Previdi, S., Ed., Filsfils, C., Ed., Decraene, B., Litkowski, S., Horneffer, M., and R. Shakir, "Source Packet Routing in Networking (SPRING) Problem Statement and Requirements", RFC 7855, DOI 10.17487/RFC7855, May 2016, <https://www.rfc-editor.org/info/rfc7855>.

Authors' Addresses

Stefano Previdi (editor)
Huawei
IT

Email: stefano@previdi.net

Les Ginsberg (editor)
Cisco Systems, Inc.
USA

Email: ginsberg@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
BE

Email: cfilsfil@cisco.com

Ahmed Bashandy
Arrcus

Email: abashandy.ietf@gmail.com

Hannes Gredler
RtBrick Inc.

Email: hannes@rtbrick.com

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2018

N. Shen, Ed.
E. Chen
A. Lindem
Cisco Systems

October 18, 2017

Carrying Geo Coordinates Information In IS-IS
draft-shen-isis-geo-coordinates-04

Abstract

This document defines a new IS-IS TLV which carries the Geo Coordinates information of the system. The Geo Coordinates information can be used by IS-IS routing or by an application.

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 http://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 21, 2018.

Copyright Notice

Table of Contents

1. Introduction

The IS-IS routing protocol defined by [ISO10589] has been widely deployed. The Geo Coordinates information can be useful, particularly within the wide area networks for numerous applications. Similar to the Dynamic Hostname defined in [RFC5301], the Geo Coordinates can also be used for network management purposes.

The Geo coordinate information can be retrieve using a variety of means (e.g., SNMP, CLI) without requiring advertising it in an IGP. Nevertheless, announcing the information in IGP allows for new applications and use cases that are elaborated hereafter.

The following provides a non-exhaustive list of sample use cases.

In the case of IGP point-to-multiple operations [I-D.lamparter-isis-p2mp], [RFC6845], the local system configuration can be greatly simplified if the outbound metric to remote neighbors can be generated automatically based on the Geo Location of the IGP neighbors.

In the application where IS-IS neighbors are on the same "sub-net", but over the WAN network, the Geo Location information may be used for equal-cost or unequal-cost load sharing on the local system. This enables location based operation on anycast IP prefixes and DMZ gateways across the WAN environment.

For the traffic matrix using the Geo Coordinates within the routing domain, instead of a collection of IP nexthops which might be translated into locations, this enables automatic region to region traffic pattern aggregation. In particular, introducing new nodes or withdrawing existing ones will be automatically reflected by the application responsible for region to region traffic aggregation. Advanced traffic engineering policies may also be enforced to avoid some nodes located on a specific region under some conditions. Such advanced TE policies are not discussed in this document.

This document describes the IS-IS protocol extension for carrying the Geo Coordinates information. A new TLV is defined for this purpose. This TLV can be distributed within the node's LSP or inside the IIH PDU. The exact mechanism an application uses the information carried in this TLV is outside the scope of this document.

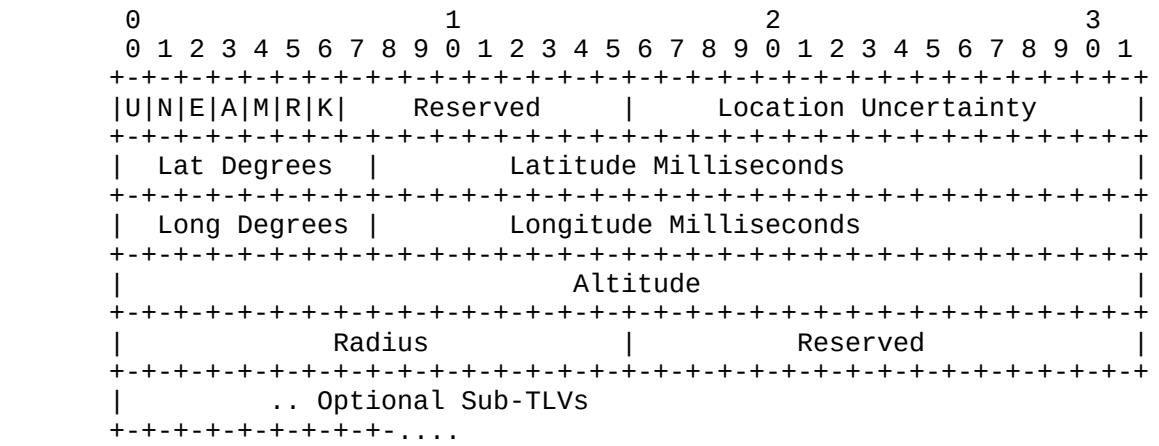
Further, it is out of scope of this document to specify how a node is provided with the information to be included in the TLV. This document does not assume whether the information included in the TLV is static or not. This is deployment-specific. Typically, this information can be used within a mobile network (trains, for example) that is grafted to a global network.

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. Packet Encoding

This Geo Coordinates extension introduces one TLV for IS-IS LSP PDU and for Hello (IIH) PDU. The code of the TLV is described in Section 4. The fields specify the location of the system using WGS-84 (World Geodetic System) reference coordinate system [WGS84]. The value of the Geo Coordinates TLV consists of the following fields:



Type: TBD. 8 bits value, to be assigned by IANA.

Length: Variable. 8 bits value. The mandatory part is 16 octets.

U-bit: If the U-bit is set, it indicates that the "Location Uncertainty" field is specified. If the U-bit is clear, it indicates the "Location Uncertainty" field is unspecified.

N-bit: If the N-bit is set, it indicates the Latitude is north relative to the Equator. If the N-bit is clear, it indicates the Latitude is south of the Equator.

E-bit: If the E-bit is set, it indicates the Longitude is east of the Prime Meridian. If the E-bit is clear, it indicates the Longitude is west of the Prime Meridian.

A-bit: If the A-bit is set, it indicates the "Altitude" field is specified. If the A-bit is clear, it indicates the "Altitude" field is unspecified.

M-bit: If the M-bit is set, it indicates the "Altitude" is specified in meters. If the M-bit is clear, it indicates the "Altitude" is in centimeters.

R-bit: If the R-bit is set, it indicates the "Radius" field is specified and the encoding is for a circular area. If the R-bit is clear, it indicates the "Radius" field is unspecified and the encoding is for a single point.

K-bit: If the K-bit is set, it indicates the "Radius" is specified in kilometers. If the K-bit is clear, it indicates the "Radius" is in meters.

Reserved: These bits are reserved. They SHOULD be set to 0 when sending protocol packets and MUST be ignored when receiving protocol packets.

Location Uncertainty: Unsigned 16-bit integer indicating the number of centimeters of uncertainty for the location.

Latitude Degrees: Unsigned 8-bit integer with a range of 0 - 90 degrees north or south of the Equator (northern or southern hemisphere, respectively).

Latitude Milliseconds: Unsigned 24-bit integer with a range of 0 - 3,599,999 (i.e., less than 60 minutes).

Longitude Degrees: Unsigned 8-bit integer with a range of 0 - 180 degrees east or west of the Prime Meridian.

Longitude Milliseconds: Unsigned 24-bit integer with a range of 0 - 3,599,999 (i.e., less than 60 minutes).

Altitude: Signed 32-bit integer containing the Height relative to sea level in centimeters or meters. A negative height indicates that the location is below sea level.

Radius: Unsigned 16-bit integer containing the radius of a circle centered at the specified coordinates. The radius is specified in meters unless the K-bit is specified indicating specification in kilometers. If the radius is specified, the geo-coordinates specify the entire area of the circle defined by the radius and center point. While the use cases herein do not make use of this field, future use cases may.

Optional Sub-TLV: Not defined in this document, for future extension related to the Geo Coordinates information.

3. Operations

The IS-IS Geo Coordinates TLV may be included in the node’s LSP, and it is recommended to be in the LSP fragment zero. This TLV can also be optionally included in the IIH PDU. This can be useful when the application is setting the outbound p2mp circuit metric based on the neighbor’s location. This can also be used in the Spine-Leaf extension [I-D.shen-isis-spine-leaf-ext] where there is no LSP being flooded into the leaf nodes.

The Geo location information can be provisioned on the system, or it can be dynamically acquired from the GPS capable device on the system.

Further, this specification assumes that the Geo Location coordinates MUST NOT be included by default. An explicit configuration parameter is required to instruct an IS-IS node to include this TLV in its announcement. If a node is instructed to include the TLV, but no value is provided, the TLV MUST NOT be announced.

4. IANA Considerations

A new TLV codepoint is defined in this document and needs to be assigned by IANA from the "IS-IS TLV Codepoints" registry. It is referred to as the Geo Coordinates TLV. This TLV is only to be optionally inserted in the LSP PDU and the IIH PDU. This document does not propose any sub-TLV out of this Geo Coordinates TLV.

Value	Name	IIH	LSP	SNP	Purge
TBD	Geo Coordinates	y	y	n	n

5. Security Considerations

Since the Geo Location coordinates may provide the exact location of the routing devices, disclosure may make the IS-IS devices more susceptible to physical attacks if such IS-IS messages are advertised outside an administrative domain. In situations where this is a concern (e.g., in military applications, or the topology of the network is considered proprietary information), the implementation MUST allow the Geo Location extension to be removed from the IS-IS advertisement. As mentioned in Section 3, the TLV is not included by default. Doing so, allow to avoid misuses of the TLV in the contexts that are not requiring such TLV to be advertised.

Security concerns for the base IS-IS are addressed in [ISO10589], [RFC5304], [RFC5310], and [RFC7602].

6. Privacy Considerations

If the location of an IS-IS router advertising Geo Location coordinates as described herein can be directly correlated to an individual, individuals, or an organization, the location of that router should be considered sensitive and IS-IS LSP containing such geo coordinates should be advertised confidentially as described in Section 5. Additionally, IS-IS network management facilities may require added authorization to view the contents of IS-IS LSPs containing geo-Location TLVs. Refer to [RFC6973] for more information.

The Uncertainty and Confidence metrics for geo-location information as described in [RFC7459] are not included in the Geo Coordinates

TLV. In a future document, these may be considered for inclusion with additional Geo Location Sub-TLVs dependent on both on requirements and adoption of [RFC7459].

7. Acknowledgments

The encoding of the Geo location is adapted from the "Geo Coordinate LISP Canonical Address Format" specified in the "LISP Canonical Address Format (LCAF)". We would like to thank the authors of that Document and particularly Dino Farinacci for subsequent discussions.

Thanks to Mohamed Boucadair, Les Ginsberg, Yi Yang, and Joe Hildebrand for commenting and discussions of Geo Coordinates precision encoding. Thanks to David Ward for commenting on attack vector in relation to this new capability of IS-IS.

8. Document Change Log

8.1. Changes to draft-shen-isis-geo-coordinates-04.txt

- Clarification and more precise descriptions throughout the document thanks to the detailed comments from Mohamed Boucadair.

8.2. Changes to draft-shen-isis-geo-coordinates-03.txt

- The 03 version submitted in April 2017 without content change.

8.3. Changes to draft-shen-isis-geo-coordinates-02.txt

- The 02 version submitted in October 2016.
- Changed the format of Geo Location encoding to have Radius field and flags to be compatible with LISP [LISP-GEO].
- Added the privacy section.

8.4. Changes to draft-shen-isis-geo-coordinates-01.txt

- The 01 version submitted in February 2016.
- Change Geo Location encoding to have better precision and to include uncertainty information.
- Added the discussion in security section for the awareness of increased probability in attack vector.

8.5. Changes to draft-shen-isis-geo-coordinates-00.txt

- o Initial version of the draft is published in February 2016.

9. References

9.1. Normative References

[ISO10589] ISO "International Organization for Standardization", "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

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

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

[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>.

[RFC6845] Sheth, N., Wang, L., and J. Zhang, "OSPF Hybrid Broadcast and Point-to-Multipoint Interface Type", RFC 6845, DOI 10.17487/RFC6845, January 2013, <https://www.rfc-editor.org/info/rfc6845>.

[RFC7602] Chunduri, U., Lu, W., Tian, A., and N. Shen, "IS-IS Extended Sequence Number TLV", RFC 7602, DOI 10.17487/RFC7602, July 2015, <https://www.rfc-editor.org/info/rfc7602>.

9.2. Informative References

[I-D.lamparter-isis-p2mp]
Franke, C., Lamparter, D., and C. Hopps, "IS-IS Point-to-Multipoint operation", draft-lamparter-isis-p2mp-01 (work in progress), October 2015.

[I-D.shen-isis-spine-leaf-ext]
Shen, N., Ginsberg, L., and S. Thyamagundalu, "IS-IS Routing for Spine-Leaf Topology", draft-shen-isis-spine-leaf-ext-03 (work in progress), March 2017.

[LISP-GEO]
Farinacci, D., "LISP Geo-Coordinate Use-Cases", draft-farinacci-lisp-geo-02 (work in progress), 2016.

[RFC6973] Cooper, A., Tschofenig, H., Aboba, B., Peterson, J., Morris, J., Hansen, M., and R. Smith, "Privacy Considerations for Internet Protocols", RFC 6973, DOI 10.17487/RFC6973, July 2013, <https://www.rfc-editor.org/info/rfc6973>.

[RFC7459] Thomson, M. and J. Winterbottom, "Representation of Uncertainty and Confidence in the Presence Information Data Format Location Object (PIDF-LO)", RFC 7459, DOI 10.17487/RFC7459, February 2015, <https://www.rfc-editor.org/info/rfc7459>.

[WGS84] National Imagery and Mapping Agency, "Department of Defense World Geodetic System 1984, Third Edition", NIMA TR8350.2, January 2000.

Authors' Addresses

Naiming Shen (editor)
Cisco Systems
560 McCarthy Blvd.
Milpitas, CA 95035
US

Email: naiming@cisco.com

Enke Chen
Cisco Systems
560 McCarthy Blvd.
Milpitas, CA 95035
US

Email: enkechen@cisco.com

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

Email: acee@cisco.com

Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2019

N. Shen
L. Ginsberg
Cisco Systems
S. Thyamagundalu

October 16, 2018

IS-IS Routing for Spine-Leaf Topology
draft-shen-isis-spine-leaf-ext-07

Abstract

This document describes a mechanism for routers and switches in a Spine-Leaf type topology to have non-reciprocal Intermediate System to Intermediate System (IS-IS) routing relationships between the leafs and spines. The leaf nodes do not need to have the topology information of other nodes and exact prefixes in the network. This extension also has application in the Internet of Things (IoT).

Table of Contents

1. Introduction

The IS-IS routing protocol defined by [ISO10589] has been widely deployed in provider networks, data centers and enterprise campus environments. In the data center and enterprise switching networks, a Spine-Leaf topology is commonly used. This document describes a mechanism where IS-IS routing can be optimized for a Spine-Leaf topology.

In a Spine-Leaf topology, normally a leaf node connects to a number of spine nodes. Data traffic going from one leaf node to another leaf node needs to pass through one of the spine nodes. Also, the decision to choose one of the spine nodes is usually part of equal cost multi-path (ECMP) load sharing. The spine nodes can be considered as gateway devices to reach destinations on other leaf nodes. In this type of topology, the spine nodes have to know the topology and routing information of the entire network, but the leaf nodes only need to know how to reach the gateway devices to which are the spine nodes they are uplinked.

This document describes the IS-IS Spine-Leaf extension that allows the spine nodes to have all the topology and routing information, while keeping the leaf nodes free of topology information other than the default gateway routing information. The leaf nodes do not even need to run a Shortest Path First (SPF) calculation since they have no topology information.

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. Motivations

- The leaf nodes in a Spine-Leaf topology do not require complete topology and routing information of the entire domain since their forwarding decision is to use ECMP with spine nodes as default gateways

- The spine nodes in a Spine-Leaf topology are richly connected to leaf nodes, which introduces significant flooding duplication if they flood all Link State PDUs (LSPs) to all the leaf nodes. It saves both spine and leaf nodes' CPU and link bandwidth resources if flooding is blocked to leaf nodes. For small Top of the Rack (ToR) leaf switches in data centers, it is meaningful to prevent full topology routing information and massive database flooding through those devices.

- When a spine node advertises a topology change, every leaf node connected to it will flood the update to all the other spine nodes, and those spine nodes will further flood them to all the leaf nodes, causing a O(n^2) flooding storm which is largely redundant.

- Similar to some of the overlay technologies which are popular in data centers, the edge devices (leaf nodes) may not need to contain all the routing and forwarding information on the device's control and forwarding planes. "Conversational Learning" can be utilized to get the specific routing and forwarding information in the case of pure CLOS topology and in the events of link and node down.

- Small devices and appliances of Internet of Things (IoT) can be considered as leafs in the routing topology sense. They have CPU and memory constrains in design, and those IoT devices do not have to know the exact network topology and prefixes as long as there are ways to reach the cloud servers or other devices.

3. Spine-Leaf (SL) Extension

3.1. Topology Examples

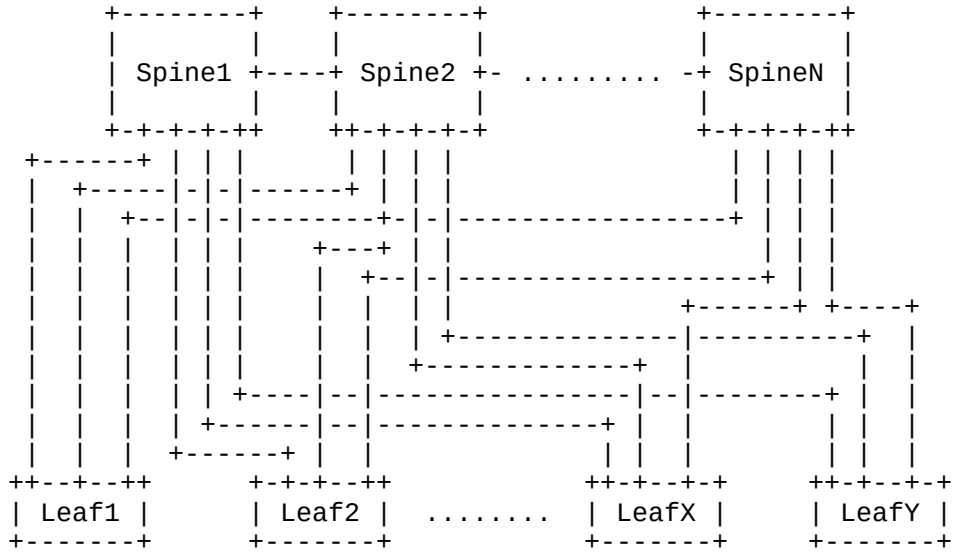


Figure 1: A Spine-Leaf Topology

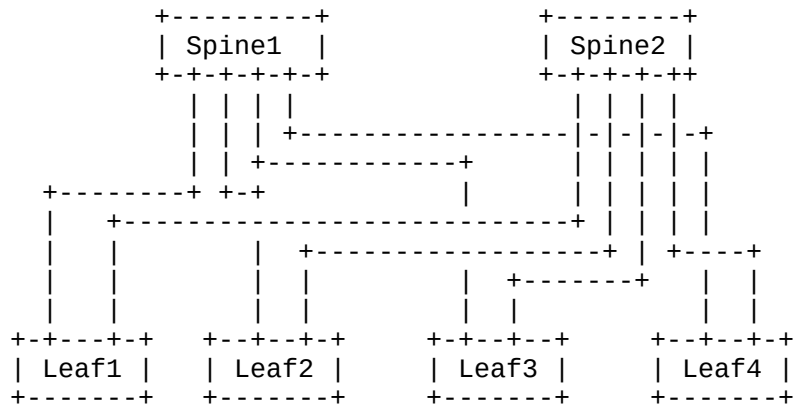


Figure 2: A CLOS Topology

3.2. Applicability Statement

This extension assumes the network is a Spine-Leaf topology, and it should not be applied in an arbitrary network setup. The spine nodes can be viewed as the aggregation layer of the network, and the leaf nodes as the access layer of the network. The leaf nodes use a load sharing algorithm with spine nodes as nexthops in routing and forwarding.

This extension works when the spine nodes are inter-connected, and it works with a pure CLOS or Fat Tree topology based network where the spines are NOT horizontally interconnected.

Although the example diagram in Figure 1 shows a fully meshed Spine-Leaf topology, this extension also works in the case where they are partially meshed. For instance, leaf1 through leaf10 may be fully meshed with spine1 through spine5 while leaf11 through leaf20 is fully meshed with spine4 through spine8, and all the spines are inter-connected in a redundant fashion.

This extension can also work in multi-level spine-leaf topology. The lower level spine node can be a 'leaf' node to the upper level spine node. A spine-leaf 'Tier' can be exchanged with IS-IS hello packets to allow tier X to be connected with tier X+1 using this extension. Normally tier-0 will be the TOR routers and switches if provisioned.

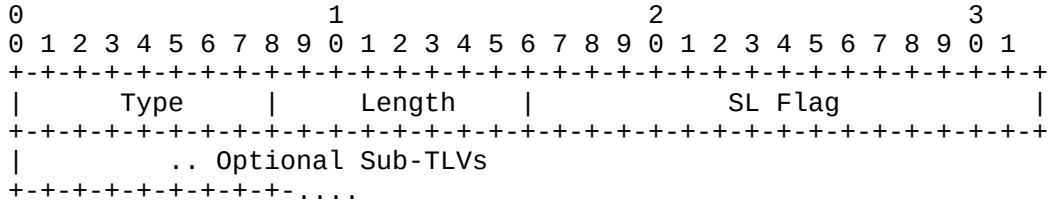
This extension also works with normal IS-IS routing in a topology with more than two layers of spine and leaf. For instance, in example diagrams Figure 1 and Figure 2, there can be another Core layer of routers/switches on top of the aggregation layer. From an IS-IS routing point of view, the Core nodes are not affected by this

extension and will have the complete topology and routing information just like the spine nodes. To make the network even more scalable, the Core layer can operate as a level-2 IS-IS sub-domain while the Spine and Leaf layers operate as stays at the level-1 IS-IS domain.

This extension assumes the link between the spine and leaf nodes are point-to-point, or point-to-point over LAN [RFC5309]. The links connecting among the spine nodes or the links between the leaf nodes can be any type.

3.3. Spine-Leaf TLV

This extension introduces a new TLV, the Spine-Leaf TLV, which may be advertised in IS-IS Hello (IIH) PDUs, LSPs, or in Circuit Scoped Link State PDUs (CS-LSP) [RFC7356]. It is used by both spine and leaf nodes in this Spine-Leaf mechanism.

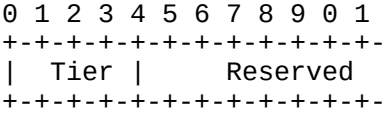


The fields of this TLV are defined as follows:

Type: 1 octet Suggested value 150 (to be assigned by IANA)

Length: 1 octet (2 + length of sub-TLVs).

SL Flags: 16 bits



Tier: A value from 0 to 15. It represents the spine-leaf tier level. The value 15 is reserved to indicate the tier level is unknown. This value is only valid when the 'T' bit (see below) is set. If the 'T' bit is clear, this value MUST be set to zero on transmission, and it MUST be ignored on receipt.

L bit (0x01): Only leaf node sets this bit. If the L bit is set in the SL flag, the node indicates it is in 'Leaf-Mode'.

R bit (0x02): Only Spine node sets this bit. If the R bit is set, the node indicates to the leaf neighbor that it can be used as the default route gateway.

T bit (0x04): If set, the value in the "Tier" field (see above) is valid.

Optional Sub-TLV: Not defined in this document, for future extension

sub-TLVs MAY be included when the TLV is in a CS-LSP. sub-TLVs MUST NOT be included when the TLV is in an IIH

3.3.1. Spine-Leaf Sub-TLVs

If the data center topology is a pure CLOS or Fat Tree, there are no link connections among the spine nodes. If we also assume there is not another Core layer on top of the aggregation layer, then the traffic from one leaf node to another may have a problem if there is a link outage between a spine node and a leaf node. For instance, in the diagram of Figure 2, if Leaf1 sends data traffic to Leaf3 through Spine1 node, and the Spine1-Leaf3 link is down, the data traffic will be dropped on the Spine1 node.

To address this issue spine and leaf nodes may send/request specific reachability information via the sub-TLVs defined below.

Two Spine-Leaf sub-TLVs are defined. The Leaf-Set sub-TLV and the Info-Req sub-TLV.

3.3.1.1. Leaf-Set Sub-TLV

This sub-TLV is used by spine nodes to optionally advertise Leaf neighbors to other Leaf nodes. The fields of this sub-TLV are defined as follows:

Type: 1 octet Suggested value 1 (to be assigned by IANA)

Length: 1 octet MUST be a multiple of 6 octets.

Leaf-Set: A list of IS-IS System-ID of the leaf node neighbors of this spine node.

3.3.1.2. Info-Req Sub-TLV

This sub-TLV is used by leaf nodes to request the advertisement of more specific prefix information from a selected spine node. The list of leaf nodes in this sub-TLV reflects the current set of leaf-nodes for which not all spine node neighbors have indicated the presence of connectivity in the Leaf-Set sub-TLV (See Section 3.3.1.1). The fields of this sub-TLV are defined as follows:

Type: 1 octet Suggested value 2 (to be assigned by IANA)

Length: 1 octet. It MUST be a multiple of 6 octets.

Info-Req: List of IS-IS System-IDs of leaf nodes for which connectivity information is being requested.

3.3.2. Advertising IPv4/IPv6 Reachability

In cases where connectivity between a leaf node and a spine node is down, the leaf node MAY request reachability information from a spine node as described in Section 3.3.1.2. The spine node utilizes TLVs 135 [RFC5305] and TLVs 236 [RFC5308] to advertise this information. These TLVs MAY be included either in IIHs or CS-LSPs [RFC7356] sent from the spine to the requesting leaf node. Sending such information in IIHs has limited scale - all reachability information MUST fit within a single IIH. It is therefore recommended that CS-LSPs be used.

3.3.3. Advertising Connection to RF-Leaf Node

For links between Spine and Leaf Nodes on which the Spine Node has set the R-bit and the Leaf node has set the L-bit in their respective Spine-Leaf TLVs, spine nodes may advertise the link with a bit in the "link-attribute" sub-TLV [RFC5029] to express this link is not used for LSP flooding. This information can be used by nodes computing a flooding topology e.g., [DYNAMIC-FLOODING], to exclude the RF-Leaf nodes from the computed flooding topology.

3.4. Mechanism

Leaf nodes in a spine-leaf application using this extension are provisioned with two attributes:

1)Tier level of 0. This indicates the node is a Leaf Node. The value 0 is advertised in the Tier field of Spine-Leaf TLV defined above.

2)Flooding reduction enabled/disabled. If flooding reduction is enabled the L-bit is set to one in the Spine-Leaf TLV defined above

A spine node does not need explicit configuration. Spine nodes can dynamically discover their tier level by computing the number of hops to a leaf node. Until a spine node determines its tier level it MUST advertise level 15 (unknown tier level) in the Spine-Leaf TLV defined above. Each tier level can also be statically provisioned on the node.

When a spine node receives an IIH which includes the Spine-Leaf TLV with Tier level 0 and 'L' bit set, it labels the point-to-point interface and adjacency to be a 'Reduced Flooding Leaf-Peer (RF-Leaf)'. IIHs sent by a spine node on a link to an RF-Leaf include the Spine-Leaf TLV with the 'R' bit set in the flags field. The 'R' bit indicates to the RF-Leaf neighbor that the spine node can be used as a default routing nexthop.

There is no change to the IS-IS adjacency bring-up mechanism for Spine-Leaf peers.

A spine node blocks LSP flooding to RF-Leaf adjacencies, except for the LSP PDUs in which the IS-IS System-ID matches the System-ID of the RF-Leaf neighbor. This exception is needed since when the leaf node reboots, the spine node needs to forward to the leaf node non-purged LSPs from the RF-Leaf's previous incarnation.

Leaf nodes will perform IS-IS LSP flooding as normal over all of its IS-IS adjacencies, but in the case of RF-Leafs only self-originated LSPs will exist in its LSP database.

Spine nodes will receive all the LSP PDUs in the network, including all the spine nodes and leaf nodes. It will perform Shortest Path First (SPF) as a normal IS-IS node does. There is no change to the route calculation and forwarding on the spine nodes.

The LSPs of a node only floods north bound towards the upper layer spine nodes. The default route is generated with loadsharing also towards the upper layer spine nodes.

RF-Leaf nodes do not have any LSP in the network except for its own. Therefore there is no need to perform SPF calculation on the RF-Leaf node. It only needs to download the default route with the nexthops of those Spine Neighbors which have the 'R' bit set in the Spine-Leaf TLV in IIH PDUs. IS-IS can perform equal cost or unequal cost load sharing while using the spine nodes as nexthops. The aggregated metric of the outbound interface and the 'Reverse Metric' [REVERSE-METRIC] can be used for this purpose.

3.4.1. Pure CLOS Topology

In a data center where the topology is pure CLOS or Fat Tree, there is no interconnection among the spine nodes, and there is not another Core layer above the aggregation layer with reachability to the leaf nodes. When flooding reduction to RF-Leafs is in use, if the link between a spine and a leaf goes down, there is then a possibility of black holing the data traffic in the network.

As in the diagram Figure 2, if the link Spine1-Leaf3 goes down, there needs to be a way for Leaf1, Leaf2 and Leaf4 to avoid the Spine1 if the destination of data traffic is to Leaf3 node.

In the above example, the Spine1 and Spine2 are provisioned to advertise the Leaf-Set sub-TLV of the Spine-Leaf TLV. Originally both Spines will advertise Leaf1 through Leaf4 as their Leaf-Set. When the Spine1-Leaf3 link is down, Spine1 will only have Leaf1, Leaf2 and Leaf4 in its Leaf-Set. This allows the other leaf nodes to know that Spine1 has lost connectivity to the leaf node of Leaf3.

Each RF-Leaf node can select another spine node to request for some prefix information associated with the lost leaf node. In this diagram of Figure 2, there are only two spine nodes (Spine-Leaf topology can have more than two spine nodes in general). Each RF-Leaf node can independently select a spine node for the leaf information. The RF-Leaf nodes will include the Info-Req sub-TLV in the Spine-Leaf TLV in hellos sent to the selected spine node, Spine2 in this case.

The spine node, upon receiving the request from one or more leaf nodes, will find the IPv6/IPv4 prefixes advertised by the leaf nodes listed in the Info-Req sub-TLV. The spine node will use the mechanism defined in Section 3.3.2 to advertise these prefixes to the RF-Leaf node. For instance, it will include the IPv4 loopback prefix of leaf3 based on the policy configured or administrative tag attached to the prefixes. When the leaf nodes receive the more specific prefixes, they will install the advertised prefixes towards the other spine nodes (Spine2 in this example).

For instance in the data center overlay scenario, when any IP destination or MAC destination uses the leaf3’s loopback as the tunnel nexthop, the overlay tunnel from leaf nodes will only select Spine2 as the gateway to reach leaf3 as long as the Spine1-Leaf3 link is still down.

In cases where multiple links or nodes fail at the same time, the RF-leaf node may need to send the Info-Req to multiple upper layer spine

nodes in order to obtain reachability information for all the partially connected nodes.

This negative routing is more useful between tier 0 and tier 1 spine-leaf levels in a multi-level spine-leaf topology when the reduced flooding extension is in use. Nodes in tiers 1 or greater may have much richer topology information and alternative paths.

3.5. Implementation and Operation

3.5.1. CSNP PDU

In Spine-Leaf extension, Complete Sequence Number PDU (CSNP) does not need to be transmitted over the Spine-Leaf link to an RF-Leaf. Some IS-IS implementations send periodic CSNPs after the initial adjacency bring-up over a point-to-point interface. There is no need for this optimization here since the RF-Leaf does not need to receive any other LSPs from the network, and the only LSPs transmitted across the Spine-Leaf link is the leaf node LSP.

Also in the graceful restart case[RFC5306], for the same reason, there is no need to send the CSNPs over the Spine-Leaf interface to an RF-Leaf. Spine nodes only need to set the SRMflag on the LSPs belonging to the RF-Leaf.

3.5.2. Overload Bit

The leaf node SHOULD set the ’overload’ bit on its LSP PDU, since if the spine nodes were to forward traffic not meant for the local node, the leaf node does not have the topology information to prevent a routing/forwarding loop.

3.5.3. Spine Node Hostname

This extension creates a non-reciprocal relationship between the spine node and leaf node. The spine node will receive leaf’s LSP and will know the leaf’s hostname, but the leaf does not have spine’s LSP. This extension allows the Dynamic Hostname TLV [RFC5301] to be optionally included in spine’s IIH PDU when sending to a ’Leaf-Peer’. This is useful in troubleshooting cases.

3.5.4. IS-IS Reverse Metric

This metric is part of the aggregated metric for leaf’s default route installation with load sharing among the spine nodes. When a spine node is in ’overload’ condition, it should use the IS-IS Reverse Metric TLV in IIH [REVERSE-METRIC] to set this metric to maximum to discourage the leaf using it as part of the loadsharing.

In some cases, certain spine nodes may have less bandwidth in link provisioning or in real-time condition, and it can use this metric to signal to the leaf nodes dynamically.

In other cases, such as when the spine node loses a link to a particular leaf node, although it can redirect the traffic to other spine nodes to reach that destination leaf node, but it MAY want to increase this metric value if the inter-spine connection becomes over utilized, or the latency becomes an issue.

In the leaf-leaf link as a backup gateway use case, the 'Reverse Metric' SHOULD always be set to very high value.

3.5.5. Spine-Leaf Traffic Engineering

Besides using the IS-IS Reverse Metric by the spine nodes to affect the traffic pattern for leaf default gateway towards multiple spine nodes, the IPv6/IPv4 Info-Advertise sub-TLVs can be selectively used by traffic engineering controllers to move data traffic around the data center fabric to alleviate congestion and to reduce the latency of a certain class of traffic pairs. By injecting more specific leaf node prefixes, it will allow the spine nodes to attract more traffic on some underutilized links.

3.5.6. Other End-to-End Services

Losing the topology information will have an impact on some of the end-to-end network services, for instance, MPLS TE or end-to-end segment routing. Some other mechanisms such as those described in PCE [RFC4655] based solution may be used. In this Spine-Leaf extension, the role of the leaf node is not too much different from the multi-level IS-IS routing while the level-1 IS-IS nodes only have the default route information towards the node which has the Attach Bit (ATT) set, and the level-2 backbone does not have any topology information of the level-1 areas. The exact mechanism to enable certain end-to-end network services in Spine-Leaf network is outside the scope of this document.

3.5.7. Address Family and Topology

IPv6 Address families[RFC5308], Multi-Topology (MT)[RFC5120] and Multi-Instance (MI)[RFC8202] information is carried over the IIH PDU. Since the goal is to simplify the operation of IS-IS network, for the simplicity of this extension, the Spine-Leaf mechanism is applied the same way to all the address families, MTs and MIs.

3.5.8. Migration

For this extension to be deployed in existing networks, a simple migration scheme is needed. To support any leaf node in the network, all the involved spine nodes have to be upgraded first. So the first step is to migrate all the involved spine nodes to support this extension, then the leaf nodes can be enabled with 'Leaf-Mode' one by one. No flag day is needed for the extension migration.

4. IANA Considerations

A new TLV codepoint is defined in this document and needs to be assigned by IANA from the "IS-IS TLV Codepoints" registry. It is referred to as the Spine-Leaf TLV and the suggested value is 150. This TLV is only to be optionally inserted either in the IIH PDU or in the Circuit Flooding Scoped LSP PDU. IANA is also requested to maintain the SL-flag bit values in this TLV, and 0x01, 0x02 and 0x04 bits are defined in this document.

Value	Name	IIH	LSP	SNP	Purge	CS-LSP
150	Spine-Leaf	y	y	n	n	y

This extension also proposes to have the Dynamic Hostname TLV, already assigned as code 137, to be allowed in IIH PDU.

Value	Name	IIH	LSP	SNP	Purge
137	Dynamic Name	y	y	n	y

Two new sub-TLVs are defined in this document and needs to be added assigned by IANA from the "IS-IS TLV Codepoints". They are referred to in this document as the Leaf-Set sub-TLV and the Info-Req sub-TLV. It is suggested to have the values 1 and 2 respectively.

This document also requests that IANA allocate from the registry of link-attribute bit values for sub-TLV 19 of TLV 22 (Extended IS reachability TLV). This new bit is referred to as the "Connect to RF-Leaf Node" bit.

Value	Name	Reference
0x3	Connect to RF-Leaf Node	This document

5. Security Considerations

Security concerns for IS-IS are addressed in [ISO10589], [RFC5304], [RFC5310], and [RFC7602]. This extension does not raise additional security issues.

6. Acknowledgments

The authors would like to thank Tony Przygienda for his discussion and contributions. The authors also would like to thank Acee Lindem, Russ White and Christian Hopps for their review and comments of this document.

7. Document Change Log

7.1. Changes to draft-shen-isis-spine-leaf-ext-05.txt

- Submitted January 2018.
- Just a refresh.

7.2. Changes to draft-shen-isis-spine-leaf-ext-04.txt

- Submitted June 2017.
- Added the Tier level information to handle the multi-level spine-leaf topology using this extension.

7.3. Changes to draft-shen-isis-spine-leaf-ext-03.txt

- Submitted March 2017.
- Added the Spine-Leaf sub-TLVs to handle the case of data center pure CLOS topology and mechanism.
- Added the Spine-Leaf TLV and sub-TLVs can be optionally inserted in either IIH PDU or CS-LSP PDU.
- Allow use of prefix Reachability TLVs 135 and 236 in IIHs/CS-LSPs sent from spine to leaf.

7.4. Changes to draft-shen-isis-spine-leaf-ext-02.txt

- Submitted October 2016.
- Removed the ’Default Route Metric’ field in the Spine-Leaf TLV and changed to using the IS-IS Reverse Metric in IIH.

7.5. Changes to draft-shen-isis-spine-leaf-ext-01.txt

- Submitted April 2016.
- No change. Refresh the draft version.

7.6. Changes to draft-shen-isis-spine-leaf-ext-00.txt

- Initial version of the draft is published in November 2015.

8. References

8.1. Normative References

[ISO10589] ISO "International Organization for Standardization", "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

[REVERSE-METRIC] Shen, N., Amante, S., and M. Abrahamsson, "IS-IS Routing with Reverse Metric", draft-ietf-isis-reverse-metric-07 (work in progress), 2017.

[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>.

[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>.

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

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

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

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

[RFC5306] Shand, M. and L. Ginsberg, "Restart Signaling for IS-IS", RFC 5306, DOI 10.17487/RFC5306, October 2008, <https://www.rfc-editor.org/info/rfc5306>.

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

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

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

[RFC7602] Chunduri, U., Lu, W., Tian, A., and N. Shen, "IS-IS Extended Sequence Number TLV", RFC 7602, DOI 10.17487/RFC7602, July 2015, <https://www.rfc-editor.org/info/rfc7602>.

[RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <https://www.rfc-editor.org/info/rfc8202>.

8.2. Informative References

[DYNAMIC-FLOODING] Li, T., "Dynamic Flooding on Dense Graphs", draft-li-dynamic-flooding (work in progress), 2018.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

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

Authors' Addresses

Naiming Shen
Cisco Systems
560 McCarthy Blvd.
Milpitas, CA 95035
US

Email: naiming@cisco.com

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

Email: ginsberg@cisco.com

Sanjay Thyamagundalu

Email: tsanjay@gmail.com

IS-IS Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 30, 2017

J. Tantsura
U. Chunduri
Individual

September 26, 2016

Signaling MSD (Maximum SID Depth) using IS-IS
draft-tantsura-isis-segment-routing-msd-02

Abstract

This document proposes a way to expose Maximum SID Depth (MSD) supported by a node at node and/or link level by an ISIS Router. In a Segment Routing (SR) enabled network a centralized controller that programs SR tunnels at the head-end node needs to know the MSD information at node level and/or link level to push the label stack of an appropriate depth.

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 http://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 March 30, 2017.

Copyright Notice

Table of Contents

1. Introduction

When Segment Routing tunnels are computed by a centralized controller, it is crucial that the controller knows the MSD "Maximum SID Depth" of the node or link SR tunnel exits over, so it doesn't download a path with SID (label stack) of a depth more than the node or link used is capable of imposing. This document describes how to use IS-IS to expose the MSD of the node or link to a centralized controller.

PCEP SR extensions [I-D.ietf-pce-segment-routing] has defined MSD, to signal in SR PCE Capability TLV, METRIC Object. However, If PCEP is not supported by a node (head-end of the SR tunnel) and controller does not participate in IGP routing it has no way to learn the MSD of the node or link configured. BGP-LS [RFC7752] defines a way to expose topology and associated different attributes, capabilities of the nodes in that topology to a centralized controller and MSD has been defined in [I-D.tantsura-idr-bgp-ls-segment-routing-msd]. For this information to be advertised by BGP for the all nodes and links of the network, where this is provisioned, IS-IS module should have this information in the LSDB.

[I-D.ietf-isis-mpls-elc] defines, RLSDC which indicates how many labels a node can read to take a decision to insert an Entropy Label (EL) and is different than how many labels a node can push as defined by MSD in this draft.

1.1. Conventions used in this document

1.1.1. Terminology

BGP-LS: Distribution of Link-State and TE Information using Border Gateway Protocol

ISIS: Intermediate System to Intermediate System

MSD: Maximum SID Depth

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol

SID: Segment Identifier

SR: Segment Routing

1.2. 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. Terminology

This memo makes use of the terms defined in [RFC4971].

3. Node MSD Advertisement

A new sub-TLV within the body of IS-IS Router Capability TLV [RFC4971], called Node MSD sub-TLV is defined to carry the provisioned SID depth of the router originating the Router Capability TLV. Node MSD is the lowest MSD supported by the node and can be provisioned in IS-IS instance.

The Type (1 byte) of this sub-TLV is TBD.

Length is 1 bytes, and

the Value field contains MSD of the router originating the Router Capability TLV. Node MSD is a number in the range of 0-254. 0 represents lack of the ability to push MSD of any depth; any other value represents that of the node. This value SHOULD represent the lowest value supported by node.

This TLV is optional. The scope of the advertisement is specific to the deployment.

4. LINK MSD Advertisement

A new sub-TLV called Link MSD sub-TLV is defined to carry the provisioned SID depth of the interface associated with the link.

The Type (1 byte) of this TLV is TBD.

Length is 1 byte, and

the Value field contains Link MSD of the router originating the corresponding IS extended reachability TLV [RFC5305] or MT IS TLV [RFC5120]. Link MSD is a number in the range of 0-254. 0 represents lack of the ability to push MSD of any depth; any other value represents that of the particular link MSD value.

5. Acknowledgements

TBD

6. IANA Considerations

This document includes a request to IANA to allocate sub-TLV type codes for the new TLV proposed in Section 3 of this document from IS-IS Router Capability TLV Registry as defined by [RFC4971]. Also for link MSD, we request IANA to allocate new sub-TLV codes as defined in Section 4 from IS extended reachability TLV (22) and MT IS TLV (222) registry.

7. Security Considerations

This document describes a mechanism for advertising Segment Routing SID depth supported at node and link level information through IS-IS LSPs and does not introduce any new security issues.

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, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4971] Vasseur, JP., Ed., Shen, N., Ed., and R. Aggarwal, Ed., "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, DOI 10.17487/RFC4971, July 2007, <http://www.rfc-editor.org/info/rfc4971>.

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

8.2. Informative References

[I-D.ietf-isis-mpls-elc] Xu, X., Kini, S., Sivabalan, S., Filsfils, C., and S. Litkowski, "Signaling Entropy Label Capability Using IS-IS", draft-ietf-isis-mpls-elc-01 (work in progress), November 2015.

[I-D.ietf-pce-segment-routing] Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., Lopez, V., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-07 (work in progress), March 2016.

[I-D.tantsura-idr-bgp-ls-segment-routing-msd] Tantsura, J., Mirsky, G., Sivabalan, S., and U. Chunduri, "Signaling Maximum SID Depth using Border Gateway Protocol Link-State", draft-tantsura-idr-bgp-ls-segment-routing-msd-01 (work in progress), July 2016.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <http://www.rfc-editor.org/info/rfc1195>.

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

[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, <http://www.rfc-editor.org/info/rfc7752>.

Authors' Addresses

Jeff Tantsura
Individual

Email: jefftant.ietf@gmail.com

Uma Chunduri
Individual

Email: uma.chunduri@gmail.com