

ISIS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2018

C. Bowers
S. Hegde
Juniper Networks
July 3, 2017

Extensions to IS-IS to Associate TE Attributes with TE Attribute Sets
and SRLGs with SRLG Sets
draft-bowers-isis-te-attribute-set-00

Abstract

This document specifies encodings that allow traffic engineering attributes to be associated with different TE attribute set identifiers and SRLGs to be associated with SRLG set identifiers.

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 January 4, 2018.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Basic assumptions	2
2.1. Minimize disruption	2
2.2. Maximize flexibility	3
3. TE Link Attributes that would benefit from this new functionality	3
4. Link Attribute Set sub-TLV	4
5. TE Attribute Set usage	5
6. SRLG Set Scoped SRLG TLV	6
7. SRLG Set usage	7
8. IANA Considerations	9
9. Management Considerations	9
10. Security Considerations	9
11. Acknowledgements	9
12. References	10
12.1. Normative References	10
12.2. Informative References	10
Authors' Addresses	10

1. Introduction

IS-IS encodings allow traffic engineering (TE) attributes, such as bandwidth-related parameters and administrative groups, as well as shared risk link groups (SRLGs) to be associated with different links in the network topology. Different applications use these attributes to decide how traffic should be directed across the network.

It can be useful for different applications to use different sets of TE attributes and SRLGs to decide how traffic is directed across the network. Existing IS-IS encodings only allow for one set of TE attributes and SRLGs to be associated with a given link. This document specifies encodings that allow different sets of TE attributes and SRLGs to be associated with a given link.

2. Basic assumptions

There are several different approaches that one could take to enable this functionality. The approach taken by this document is based on the following assumptions about the use of this encoding.

2.1. Minimize disruption

The requirements of many current and future deployments of SR and RSVP can be satisfied using the existing encodings that support a single set of TE attributes and SRLGs. The encodings described here allow the advertisement of multiple sets of TE attributes and SRLGs.

They do so in a way that minimizes the disruption and burden, in terms of interoperability testing, software upgrades, and overall complexity, on deployments that do not need this more complex functionality.

2.2. Maximize flexibility

Future applications are difficult to predict, especially as network operators deploy their own customized, centralized controllers. The encodings described here does not try to associate TE attributes and SRLGs with particular applications. Instead, they allow TE attributes and SRLGs to be organized into sets, using groupings that make most sense for the network operator's particular use case. The network advertises these TE attributes and SRLGs with their associated TE attribute and SRLG set identifiers, and different applications use this information as they see fit. The authors believe that this approach provides the greatest flexibility for those networks that are likely to require these more complex capabilities.

3. TE Link Attributes that would benefit from this new functionality

There are currently 36 sub-TLVs defined for TLV#22 (as well as TLVs #23, #141, #222, and #223.) We draw a distinction between two types of sub-TLVs in TLV#22. Some sub-TLVs (such as the IPv4 interface address and neighbor address sub-TLVs) are used to identify a link. In this document, we refer to these as TE link identifier sub-TLVs. Below is a complete list of the sub-TLVs in TLV#22 that we classify as TE link identifier sub-TLVs.

Type	Description
-----	-----
4	Link Local/Remote Identifiers
6	IPv4 interface address
8	IPv4 neighbor address
12	IPv6 Interface Address
13	IPv6 Neighbor Address

sub-TLVs of TLV#22 classified as TE link identifier sub-TLVs

Since TE link identifier sub-TLVs are used to identify links, it does not make sense to allow these sub-TLVs to be advertised more than once with different values for a given link.

In principle, the remaining 31 sub-TLVs currently defined for TLV#22 are candidates for enabling the advertisement of different values scoped by a TE attribute set identifier. However, for this document

we only specify this new functionality for the following subset of TE link attributes.

Type	Description
3	Administrative group (color)
9	Maximum link bandwidth
10	Maximum reservable link bandwidth
11	Unreserved bandwidth
14	Extended Administrative Group
18	TE Default metric
33	Unidirectional Link Delay
34	Min/Max Unidirectional Link Delay
35	Unidirectional Delay Variation
36	Unidirectional Link Loss
37	Unidirectional Residual Bandwidth
38	Unidirectional Available Bandwidth
39	Unidirectional Utilized Bandwidth

Figure 1: TE link attributes sub-TLVs given the ability to be advertised with different values scoped by TE attribute set identifier

The new TE Attribute Set Identifier is a 32-bit value that identifies a set of attributes. The TE Attribute Set Identifier with a value of zero is special. Existing encodings for advertising attributes that do not explicitly support the inclusion of the TE Attribute Set Identifier are now understood to implicitly advertise attributes with the TE Attribute Set Identifier set to zero. In this framework, existing implementations using the existing encodings already support the advertisement of attributes with the TE attribute set id = 0.

In order to ensure a consistent view of the attribute set scoped attributes, for encodings that explicitly support the TE Attribute Set Identifier, advertising an attribute with TE Attribute Set Identifier set to zero is not allowed.

4. Link Attribute Set sub-TLV

The Link Attribute Set sub-TLV is a new sub-TLV for TLVs 22, 23, 141, 222, and 223. It allows multiple values of a given TE link attribute to be advertised for the same link, scoped by the TE Attribute Set ID.

Type: To be assigned by IANA (suggested value 101)

Length: Number of octets in the value field (1 octet)

Value:

TE Attribute Set Identifier: A 32-bit value containing the non-zero TE Attribute Set Identifier that identifies a set of attributes. The Link Attribute Set sub-TLV MUST be ignored if the TE Attribute Set Identifier is zero. This ensures a consistent view of the attribute set scoped link attributes, where the Link Attribute sub-TLVs advertised directly in TLV#22 are now understood to be implicitly advertised with the TE Attribute Set Identifier equal to zero.

Link Attribute sub-sub-TLVs: The format of these Link Attribute sub-sub-TLVs matches the existing formats for the Link Attribute sub-TLVs defined in [RFC5305] and [RFC7810]. Each Link Attribute sub-sub-TLV advertised in a given Link Attribute Set sub-TLV is associated with the TE Attribute Set Identifier in the Link Attribute Set sub-TLV. Figure 1 shows the subset of existing Link Attributes sub-TLVs that we are specifying in this document.

5. TE Attribute Set usage

The TE attribute set uses a simple substitution semantics. We consider the TE attribute set with identifier=0 to be the default TE attribute set. An application receiving attributes in the default TE attribute set will use those default TE attributes, unless it receives attributes in the one non-default TE attribute set that it has been configured or programmed to consider.

In many network scenarios, all of the applications will only need to use a single common set of TE attributes advertised with their existing encodings. In this framework, these applications will all be using TE attribute set = 0, the default TE attribute set.

Application	Attribute set id
-----	-----
A	0 (implicit)
B	0 (implicit)
C	0 (implicit)

Scenario where all applications use a single common set of TE attributes

In some scenarios, a network operator will need to advertise different values of a given attribute for a given link. Consider a scenario where applications D, E, and F need common values for all TE link attributes, except for sub-TLV#9 (Maximum link bandwidth). Applications D and E use a common value for sub-TLV#9, while

application F needs a different value for sub-TLV#9. This scenario is supported by having each link advertise all sub-TLVs in TLV#22 as they are advertised today. These advertisements are understood to be advertised with the TE attribute set id = 0. Applications D and E only need to use these advertisements. Links also advertise sub-sub-TLV#9 in the TE Attribute Set sub-TLV with TE attribute set id = 1. Application F is configured to use attribute set id = 1. This means that application F first looks for the value of each attribute scoped for TE attribute set = 1. If it is present, application F uses that attribute set scoped value. If it is not present, application F uses the value in the default TE attribute set (id=0).

Application	Attribute set id
-----	-----
D	0 (implicit)
E	0 (implicit)
F	1

Scenario where applications need different values for some attributes

From a standardization perspective, there is not intended to be any fixed mapping between a given TE Attribute Set Identifier and a given application. A network operator wishing to advertise different attribute sets could configure the network equipment to advertise attributes with different values of the TE Attribute Set Identifier based on their objectives. The different applications (be they controller-based applications or distributed applications) would make use of the different attribute sets based on convention within that network.

6. SRLG Set Scoped SRLG TLV

A new TLV is defined to allow SRLGs to be advertised for a given link and associated with a specific SRLG set identifier. Although similar in functionality to TLV 138 (defined by [RFC5307]) and TLV 139 (defined by [RFC6119]), a single TLV provides support for IPv4, IPv6, and unnumbered identifiers for a link. Unlike TLVs 138/139 it utilizes sub-TLVs to encode the link identifiers in order to provide the flexible formatting required to support multiple link identifier types.

Type: To be assigned by IANA (suggested value 238)

Length: Number of octets in the value field (1 octet)

Value:

Neighbor System-ID + pseudo-node ID (7 octets)

SRLG Set Identifier: A 32-bit value containing the non-zero SRLG Set Identifier that identifies a set of SRLGs. The SRLG Set Scoped SRLG TLV MUST be ignored if the SRLG Set Identifier is zero. This ensures a consistent view of the SRLG set scoped link attributes, where the SRLGs advertised directly in TLV#138 and TLV#139 are now understood to be implicitly advertised with the TE Attribute Set Identifier equal to zero.

Length of sub-TLVs in octets (1 octet)

Link Identifier sub-TLVs (variable)

0 or more SRLG Values (Each value is 4 octets)

The following Link Identifier sub-TLVs are defined. The type values are only suggested values. The actual values will be assigned by IANA. However, as the formats are identical to existing sub-TLVs defined for TLVs 22, 23, 141, 222, and 223 the assignment of the suggested sub-TLV types is strongly encouraged.

Type	Description
-----	-----
4	Link Local/Remote Identifiers
6	IPv4 interface address
8	IPv4 neighbor address
12	IPv6 Interface Address
13	IPv6 Neighbor Address

At least one set of link identifiers (IPv4, IPv6, or unnumbered) MUST be present. TLVs which do not meet this requirement MUST be ignored.

Multiple TLVs for the same link MAY be advertised.

7. SRLG Set usage

The new SRLG Set Identifier is a 32-bit value that identifies a set of SRLGs. The SRLG Set Identifier with a value of zero is special. Existing encodings for advertising SRLGs that do not explicitly support the inclusion of the SRLG Set Identifier are now understood to implicitly advertise SRLGs with the SRLG Set Identifier set to zero. In this framework, existing implementations using the existing encodings already support the advertisement of SRLGs with the SRLG set id = 0.

In order to ensure a consistent view of the SRLG set scoped SRLGs, for encodings that explicitly support the SRLG Set Identifier, advertising an attribute with SRLG Set Identifier set to zero is not allowed.

The SRLG set uses additive semantics. An application receiving SRLGs scoped with different SRLG set identifiers will take the union of the SRLGs in each SRLG set that the application is programmed to take into consideration. Given the additive semantics of SRLG sets, we do not use the SRLGs with SRLG set identifier = 0 as a default value in the absence of other SRLGs with non-zero SRLG set identifier.

The following example illustrates the expected use of the advertising SRLGs scoped with SRLG set identifiers. SRLGs in networks often follow a natural grouping into sets. As a concrete example, assume that one set of SRLGs corresponds to links within a metro area (intra-city SRLGs). A second set of SRLGs corresponds to links between metro area (inter-city SRLGs). A third set of SRLGs corresponds to links between continents on undersea cables (inter-continental SRLGs). A reasonable mapping of these natural SRLG groupings to SRLG set identifier is shown below in Figure 2. The network would be configured to advertise SRLGs scoped with these SRLG set identifiers.

Natural SRLG groupings	SRLG set id
-----	-----
intra-city SRLGs	1
inter-city SRLGs	2
inter-continental SRLGs	3

Figure 2: Example mapping of natural SRLG groupings to SRLG set identifier

Assume that the network operator starts out with two applications. Application G should take into account all three groups of SRLGs as path constraints: intra-city, inter-city, and inter-continental SRLGs. Instead, application H should only take into account inter-city and inter-continental SRLGs. This can be accomplished by having application G use union of SRLG sets 1, 2, and 3, while application H uses the union of SRLG sets 2 and 3, as shown in Figure 3.

Application	SRLG set ids
-----	-----
G	1+2+3
H	2+3

Figure 3: Example usage of SRLG sets by applications

This accomplishes the goals of the network operator in a very natural way.

Now suppose that the network operator introduces a third application, application J, that should only take into account intra-city and

inter-city SRLGs. This can be accomplished without modifying any of the SRLG advertisements. The new application J need only be programmed or configured to take use the union of SRLG sets 1 and 2, as as shown in Figure 4.

Application	SRLG set ids
-----	-----
G	1+2+3
H	2+3
J	1+2

Figure 4: The simplicity of adding a new application

If application J is a centralized controller-based application, the new application can be introduced with even touching the network itself.

8. IANA Considerations

IANA is requested to create a new sub-TLV, the Link Attribute Set sub-TLV for TLVs 22, 23, 141, 222, and 223.

Type	Description	22	23	141	222	223	Ref.
-----	-----	--	--	---	---	---	-----
TBA1	Link Attribute Set sub-TLV	y	y	y	y	y	[This draft]

IANA is requested to create a new TLV, the SRLG Set Scoped SRLG TLV.

Type	Description	IIH	SNP	LSP	Purge	Ref.
-----	-----	---	---	---	---	-----
TBA2	TE Attribute Set Scoped SRLG TLV	n	n	y	n	[This draft]

9. Management Considerations

TBD

10. Security Considerations

TBD

11. Acknowledgements

The basic format for the encoding of the Link Attribute Set sub-TLV and the TE Attribute Set Scoped SRLG TLV follows the basic format of the encodings in [I-D.ginsberg-isis-te-app].

12. References

12.1. Normative References

- [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>>.
- [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>>.
- [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>>.
- [RFC7810] Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 7810, DOI 10.17487/RFC7810, May 2016, <<http://www.rfc-editor.org/info/rfc7810>>.

12.2. Informative References

- [I-D.ginsberg-isis-te-app]
Ginsberg, L., Psenak, P., Previdi, S., and W. Henderickx,
"IS-IS TE Attributes per application", draft-ginsberg-
isis-te-app-00 (work in progress), February 2017.

Authors' Addresses

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: cbowers@juniper.net

Shraddha Hegde
Juniper Networks
Embassy Business Park
Bangalore, KA 560093
India

Email: shraddha@juniper.net

Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 22, 2017

L. Ginsberg
P. Psenak
Cisco Systems
S. Previdi
Individual
W. Henderickx
Nokia
June 20, 2017

IS-IS TE Attributes per application
draft-ginsberg-isis-te-app-03.txt

Abstract

Existing traffic engineering related link attribute advertisements have been defined and are used in RSVP-TE deployments. In cases where multiple applications wish to make use of these link attributes the current advertisements do not support application specific values for a given attribute nor do they support indication of which applications are using the advertised value for a given link.

This draft introduces new link attribute advertisements which address both of these shortcomings. It also discusses backwards compatibility issues and how to minimize duplicate advertisements in the presence of routers which do not support the extensions defined in this document.

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 December 22, 2017.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Requirements Discussion	3
3. Legacy Advertisements	4
3.1. Legacy sub-TLVs	4
3.2. Legacy SRLG Advertisements	5
4. Advertising Application Specific Link Attributes	5
4.1. Application Identifier Bit Mask	5
4.2. Application Specific Link Attributes sub-TLV	7
4.3. Application Specific SRLG TLV	8
5. Deployment Considerations	9
6. Attribute Advertisements and Enablement	10
7. Interoperability, Backwards Compatibility and Migration Concerns	10
7.1. RSVP-TE only deployments	11
7.2. Multiple Applications: Common Attributes with RSVP-TE	11
7.3. Multiple Applications: All Attributes Not Shared w RSVP-TE	11
7.4. Deprecating legacy advertisements	11
8. IANA Considerations	12
9. Security Considerations	13
10. Acknowledgements	13
11. References	13
11.1. Normative References	13
11.2. Informative References	14
Authors' Addresses	14

1. Introduction

Advertisement of link attributes by the Intermediate-System-to-Intermediate-System (IS-IS) protocol in support of traffic engineering (TE) was introduced by [RFC5305] and extended by [RFC5307], [RFC6119], and [RFC7810]. Use of these extensions has been associated with deployments supporting Traffic Engineering over Multiprotocol Label Switching (MPLS) in the presence of Resource Reservation Protocol (RSVP) - more succinctly referred to as RSVP-TE.

In recent years new applications have been introduced which have use cases for many of the link attributes historically used by RSVP-TE. Such applications include Segment Routing Traffic Engineering (SRTE) and Loop Free Alternates (LFA). This has introduced ambiguity in that if a deployment includes a mix of RSVP-TE support and SRTE support (for example) it is not possible to unambiguously indicate which advertisements are to be used by RSVP-TE and which advertisements are to be used by SRTE. If the topologies are fully congruent this may not be an issue, but any incongruence leads to ambiguity.

An additional issue arises in cases where both applications are supported on a link but the link attribute values associated with each application differ. Current advertisements do not support advertising application specific values for the same attribute on a specific link.

This document defines extensions which address these issues. Also, as evolution of use cases for link attributes can be expected to continue in the years to come, this document defines a solution which is easily extensible to the introduction of new applications and new use cases.

2. Requirements Discussion

As stated previously, evolution of use cases for link attributes can be expected to continue - so any discussion of existing use cases is limited to requirements which are known at the time of this writing. However, in order to determine the functionality required beyond what already exists in IS-IS, it is only necessary to discuss use cases which justify the key points identified in the introduction - which are:

1. Support for indicating which applications are using the link attribute advertisements on a link
2. Support for advertising application specific values for the same attribute on a link

[RFC7855] discusses use cases/requirements for SR. Included among these use cases is SRTE. If both RSVP-TE and SRTE are deployed in a network, link attribute advertisements can be used by one or both of these applications. As there is no requirement for the link attributes advertised on a given link used by SRTE to be identical to the link attributes advertised on that same link used by RSVP-TE, there is a clear requirement to indicate independently which link attribute advertisements are to be used by each application.

As the number of applications which may wish to utilize link attributes may grow in the future, an additional requirement is that the extensions defined allow the association of additional applications to link attributes without altering the format of the advertisements or introducing new backwards compatibility issues.

Finally, there may still be many cases where a single attribute value can be shared among multiple applications, so the solution must minimize advertising duplicate link/attribute pairs whenever possible.

3. Legacy Advertisements

There are existing advertisements used in support of RSVP-TE. These advertisements include sub-TLVs for TLVs 22, 23, 141, 222, and 223 and TLVs for SRLG advertisement.

3.1. Legacy sub-TLVs

Sub-TLVs for TLVs 22, 23, 141, 222, and 223

Code Point/Attribute Name

3 Administrative group (color)
9 Maximum link bandwidth
10 Maximum reservable link bandwidth
11 Unreserved bandwidth
14 Extended Administrative Group
33 Unidirectional Link Delay
34 Min/Max Unidirectional Link Delay
35 Unidirectional Delay Variation
36 Unidirectional Link Loss
37 Unidirectional Residual Bandwidth
38 Unidirectional Available Bandwidth
39 Unidirectional Utilized Bandwidth

3.2. Legacy SRLG Advertisements

TLV 138 GMPLS-SRLG

Supports links identified by IPv4 addresses and unnumbered links

TLV 139 IPv6 SRLG

Supports links identified by IPv6 addresses

Note that [RFC6119] prohibits the use of TLV 139 when it is possible to use TLV 138.

4. Advertising Application Specific Link Attributes

Two new code points are defined in support of Application Specific Link Attribute Advertisements:

1) Application Specific Link Attributes sub-TLV for TLVs 22, 23, 141, 222, and 223

2) Application Specific Shared Risk Link Group (SRLG) TLV

In support of these new advertisements, an application bit mask is defined which identifies the application(s) associated with a given advertisement.

The following sections define the format of these new advertisements.

4.1. Application Identifier Bit Mask

Identification of the set of applications associated with link attribute advertisements utilizes two bit masks. One bit mask is for standard applications where the definition of each bit is defined in a new IANA controlled registry. A second bit mask is for non-standard User Defined Applications (UDAs).

The encoding defined below is used by both the Application Specific Link Attributes sub-TLV and the Application Specific SRLG TLV.

```

0 1 2 3 4 5 6 7
+---+---+---+---+---+
|   SABML+F   | 1 octet
+---+---+---+---+---+
|   UDABML+F   | 1 octet
+---+---+---+---+---+
|   SABM       | ... 0 - 127 octets
+---+---+---+---+---+

```

```
|  UDABM          ...  0 - 127 octets
+---+---+---+---+---+---+---+---+---+
```

SABML+F (1 octet)

Standard Application Bit Mask Length/Flags

```
      0 1 2 3 4 5 6 7
+---+---+---+---+---+---+---+---+---+
|L|  SA-Length  |
+---+---+---+---+---+---+---+---+---+
```

L-flag: Applications listed (both Standard and User Defined) MUST use the legacy advertisements for the corresponding link found in TLVs 22, 23, 141, 222, and 223 or TLV 138 or TLV 139 as appropriate.

SA-Length: Indicates the length in octets (0-127) of the Bit Mask for Standard Applications.

UDABML+F (1 octet)

User Defined Application Bit Mask Length/Flags

```
      0 1 2 3 4 5 6 7
+---+---+---+---+---+---+---+---+---+
|R|  UDA-Length  |
+---+---+---+---+---+---+---+---+---+
```

R: Reserved. Transmitted as 0 and ignored on receipt

UDA-Length: Indicates the length in octets (0-127) of the Bit Mask for User Defined Applications.

SABM (variable length)

Standard Application Bit Mask

(SA-Length * 8) bits

This is omitted if SA-Length is 0.

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

R-bit: RSVP-TE

S-bit: Segment Routing Traffic Engineering

F-bit: Loop Free Alternate

UDABM (variable length)
User Defined Application Bit Mask

(UDA Length * 8) bits

```

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

This is omitted if UDA-Length is 0.

NOTE: If both SA-length and UDA-Length are zero, then the attributes associated with this attribute identifier bit mask MAY be used by any Standard Application and any User Defined Application.

Standard Application Bits are defined/sent starting with Bit 0. Additional bit definitions that may be defined in the future SHOULD be assigned in ascending bit order so as to minimize the number of octets that will need to be transmitted. Undefined bits MUST be transmitted as 0 and MUST be ignored on receipt. Bits that are NOT transmitted MUST be treated as if they are set to 0 on receipt.

User Defined Application bits have no relationship to Standard Application bits and are NOT managed by IANA or any other standards body. It is recommended that bits are used starting with Bit 0 so as to minimize the number of octets required to advertise all UDAs.

4.2. Application Specific Link Attributes sub-TLV

A new sub-TLV for TLVs 22, 23, 141, 222, and 223 is defined which supports specification of the applications and application specific attribute values.

Type: 15 (suggested value - to be assigned by IANA)
Length: Variable (1 octet)
Value:

Application Bit Mask (as defined in Section 3.1)

Link Attribute sub-sub-TLVs - format matches the existing formats defined in [RFC5305] and [RFC7810]

When the L-flag is set in the Application Identifiers, all of the applications specified in the bit mask MUST use the link attribute sub-TLV advertisements listed in Section 3.1 for the corresponding link. Application specific link attribute sub-sub-TLVs for the corresponding link attributes MUST NOT be advertised for the set of applications specified in the Standard/User Application Bit Masks and all such advertisements MUST be ignored on receipt.

Multiple sub-TLVs for the same link MAY be advertised. When multiple sub-TLVs for the same link are advertised, they SHOULD advertise non-conflicting application/attribute pairs. A conflict exists when the same application is associated with two different values of the same link attribute for a given link. In cases where conflicting values for the same application/attribute/link are advertised all the conflicting values MUST be ignored.

For a given application, the setting of the L-flag MUST be the same in all sub-TLVs for a given link. In cases where this constraint is violated, the L-flag MUST be considered set for this application.

A new registry of sub-sub-TLVs is to be created by IANA which defines the link attribute sub-sub-TLV code points. A sub-sub-TLV is defined for each of the existing sub-TLVs listed in Section 3.1. Format of the sub-sub-TLVs matches the format of the corresponding legacy sub-TLV and IANA is requested to assign the legacy sub-TLV identifier to the corresponding sub-sub-TLV.

4.3. Application Specific SRLG TLV

A new TLV is defined to advertise application specific SRLGs for a given link. Although similar in functionality to TLV 138 (defined by [RFC5307]) and TLV 139 (defined by [RFC6119]), a single TLV provides support for IPv4, IPv6, and unnumbered identifiers for a link. Unlike TLVs 138/139, it utilizes sub-TLVs to encode the link identifiers in order to provide the flexible formatting required to support multiple link identifier types.

Type: 238 (Suggested value - to be assigned by IANA)
Length: Number of octets in the value field (1 octet)
Value:
 Neighbor System-ID + pseudo-node ID (7 octets)
 Application Bit Mask (as defined in Section 3.1)
 Length of sub-TLVs (1 octet)
 Link Identifier sub-TLVs (variable)
 0 or more SRLG Values (Each value is 4 octets)

The following Link Identifier sub-TLVs are defined. The type values are suggested and will be assigned by IANA - but as the formats are identical to existing sub-TLVs defined for TLVs 22, 23, 141, 222, and 223 the use of the suggested sub-TLV types is strongly encouraged.

Type	Description
4	Link Local/Remote Identifiers (see [RFC5307])
6	IPv4 interface address (see [RFC5305])
8	IPv4 neighbor address (see [RFC5305])
12	IPv6 Interface Address (see [RFC6119])
13	IPv6 Neighbor Address (see [RFC6119])

At least one set of link identifiers (IPv4, IPv6, or unnumbered) MUST be present. TLVs which do not meet this requirement MUST be ignored.

Multiple TLVs for the same link MAY be advertised.

When the L-flag is set in the Application Identifiers, SRLG values MUST NOT be included in the TLV. Any SRLG values which are advertised MUST be ignored. Based on the link identifiers advertised the corresponding legacy TLV (see Section 3.2) can be identified and the SRLG values advertised in the legacy TLV MUST be used by the set of applications specified in the Application Bit Mask.

For a given application, the setting of the L-flag MUST be the same in all TLVs for a given link. In cases where this constraint is violated, the L-flag MUST be considered set for this application.

5. Deployment Considerations

If link attributes are advertised associated with zero length application bit masks for both standard applications and user defined applications, then that set of link attributes MAY be used by any application. If support for a new application is introduced on any node in a network in the presence of such advertisements, these advertisements MAY be used by the new application. If this is not what is intended, then existing advertisements MUST be readvertised

with an explicit set of applications specified before a new application is introduced.

6. Attribute Advertisements and Enablement

This document defines extensions to support the advertisement of application specific link attributes. The presence or absence of link attribute advertisements for a given application on a link does NOT indicate the state of enablement of that application on that link. Enablement of an application on a link is controlled by other means.

For some applications, the concept of enablement is implicit. For example, SRTE implicitly is enabled on all links which are part of the Segment Routing enabled topology. Advertisement of link attributes supports constraints which may be applied when specifying an explicit path through that topology.

For other applications enablement is controlled by local configuration. For example, use of a link as an LFA can be controlled by local enablement/disablement and/or the use of administrative tags.

It is an application specific policy as to whether a given link can be used by that application even in the absence of any application specific link attributes.

7. Interoperability, Backwards Compatibility and Migration Concerns

Existing deployments of RSVP-TE utilize the legacy advertisements listed in Section 3. Routers which do not support the extensions defined in this document will only process legacy advertisements and are likely to infer that RSVP-TE is enabled on the links for which legacy advertisements exist. It is expected that deployments using the legacy advertisements will persist for a significant period of time - therefore deployments using the extensions defined in this document must be able to co-exist with use of the legacy advertisements by routers which do not support the extensions defined in this document. The following sub-sections discuss interoperability and backwards compatibility concerns for a number of deployment scenarios.

Note that in all cases the defined strategy can be employed on a per link basis.

7.1. RSVP-TE only deployments

In deployments where RSVP-TE is the only application utilizing link attribute advertisements, use of the the legacy advertisements can continue without change.

7.2. Multiple Applications: Common Attributes with RSVP-TE

In cases where multiple applications are utilizing a given link, one of the applications is RSVP-TE, and all link attributes for a given link are common to the set of applications utilizing that link, interoperability is achieved by using legacy advertisements and sending application specific advertisements with L-bit set and no link attribute values. This avoids duplication of link attribute advertisements.

7.3. Multiple Applications: All Attributes Not Shared w RSVP-TE

In cases where one or more applications other than RSVP-TE are utilizing a given link and one or more link attribute values are NOT shared with RSVP-TE, it is necessary to use application specific advertisements as defined in this document. Attributes for applications other than RSVP-TE MUST be advertised using application specific advertisements which have the L-bit clear. In cases where some link attributes are shared with RSVP-TE, this requires duplicate advertisements for those attributes.

The discussion in this section applies to cases where RSVP-TE is NOT using any advertised attributes on a link and to cases where RSVP-TE is using some link attribute advertisements on the link but some link attributes cannot be shared with RSVP-TE.

7.4. Deprecating legacy advertisements

The extensions defined in this document support RSVP-TE as one of the supported applications - so a long term goal for deployments would be to deprecate use of the legacy advertisements in support of RSVP-TE. This can be done in the following step-wise manner:

- 1) Upgrade all routers to support extensions in this document
- 2) Readvertise all legacy link attributes using application specific advertisements with L-bit clear and R-bit set.
- 3) Remove legacy advertisements

8. IANA Considerations

This document defines a new sub-TLV for TLVs 22, 23, 141, 222, and 223.

Type	Description	22	23	141	222	223
15	Application Specific Link Attributes	y	y	y	y	y

This document defines one new TLV:

Type	Description	IIH	SNP	LSP	Purge
238	Application Specific SRLG	n	n	y	n

This document requests a new IANA registry be created to control the assignment of sub-sub-TLV codepoints for the Application Specific Link Attributes sub-TLV. The suggested name of the new registry is "sub-sub-TLV code points for application link attributes". The registration procedure is "Expert Review" as defined in [RFC5226]. The following assignments are made by this document:

Type	Description
3	Administrative group (color)
9	Maximum link bandwidth
10	Maximum reservable link bandwidth
11	Unreserved bandwidth
14	Extended Administrative Group
33	Unidirectional Link Delay
34	Min/Max Unidirectional Link Delay
35	Unidirectional Delay Variation
36	Unidirectional Link Loss
37	Unidirectional Residual Bandwidth
38	Unidirectional Available Bandwidth
39	Unidirectional Utilized Bandwidth

This document requests a new IANA registry be created to control the assignment of application bit identifiers. The suggested name of the new registry is "Link Attribute Applications". The registration procedure is "Expert Review" as defined in [RFC5226]. The following assignments are made by this document:

Bit #	Name

0	RSVP-TE (R-bit)
1	Segment Routing Traffic Engineering (S-bit)
2	Loop Free Alternate (F-bit)

This document requests a new IANA registry be created to control the assignment of sub-TLV types for the application specific SRLG TLV. The suggested name of the new registry is "Sub-TLVs for TLV 238". The registration procedure is "Expert Review" as defined in [RFC5226]. The following assignments are made by this document:

Value	Description

4	Link Local/Remote Identifiers (see [RFC5307])
6	IPv4 interface address (see [RFC5305])
8	IPv4 neighbor address (see [RFC5305])
12	IPv6 Interface Address (see [RFC6119])
13	IPv6 Neighbor Address (see [RFC6119])

9. Security Considerations

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

10. Acknowledgements

The authors would like to thank John Drake and Acee Lindem for their careful review and content suggestions.

11. References

11.1. 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>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.
- [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>>.
- [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>>.
- [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>>.
- [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>>.
- [RFC7810] Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 7810, DOI 10.17487/RFC7810, May 2016, <<http://www.rfc-editor.org/info/rfc7810>>.

11.2. Informative References

- [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, <<http://www.rfc-editor.org/info/rfc7855>>.

Authors' Addresses

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

Email: ginsberg@cisco.com

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

Email: ppsenak@cisco.com

Stefano Previdi
Individual

Email: stefano@previdi.net

Wim Henderickx
Nokia
Copernicuslaan 50
Antwerp 2018 94089
Belgium

Email: wim.henderickx@nokia.com

IS-IS WG
Internet-Draft
Intended status: Standards Track
Expires: March 19, 2018

S. Hegde
C. Bowers
Juniper Networks
P. Mattes
M. Nanduri
S. Giacalone
Microsoft
I. Mohammad
Arista Networks
September 15, 2017

Advertising TE protocols in IS-IS
draft-hegde-isis-advertising-te-protocols-03

Abstract

This document defines a mechanism to indicate which traffic engineering protocols are enabled on a link in IS-IS. It does so by introducing a new traffic-engineering protocol sub-TLV for TLV-22. This document also describes mechanisms to address backward compatibility issues for implementations that have not yet been upgraded to software that understands this new sub-TLV.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on March 19, 2018.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Goals	4
2.1. Explicit and unambiguous indication of TE protocol	4
3. Solution	5
3.1. Traffic-engineering protocol sub-TLV	5
3.2. Segment Routing flag considerations	6
4. Backward compatibility	7
4.1. Scenario with upgraded RSVP-TE transit router but RSVP-TE ingress router not upgraded	7
4.2. Scenario with upgraded RSVP-TE ingress router but RSVP-TE transit router not upgraded	8
4.3. Need for a long term solution	8
5. Security Considerations	9
6. IANA Considerations	9
7. Acknowledgements	9
8. References	9
8.1. Normative References	9
8.2. Informative References	10
Authors' Addresses	10

1. Introduction

IS-IS extensions for traffic engineering are specified in [RFC5305]. [RFC5305] defines several link attributes such as administrative group, maximum link bandwidth, and shared risk link groups (SRLGs) which can be used by traffic engineering applications. Additional link attributes for traffic engineering have subsequently been defined in other documents as well. Most recently [RFC7810] defined link attributes for delay, loss, and measured bandwidth utilization.

The primary consumers of these traffic engineering link attributes have been RSVP-based applications that use the advertised link attributes to compute paths which will subsequently be signalled using RSVP-TE. However, these traffic engineering link attributes have also been used by other applications, such as IP/LDP fast-reroute using loop-free alternates as described in [RFC7916]. In the future, it is likely that traffic engineering applications based on Segment Routing [I-D.ietf-spring-segment-routing] will also use these link attributes.

Existing IS-IS standards do not provide a mechanism to explicitly indicate whether or not RSVP has been enabled on a link. Instead, different RSVP-TE implementations have used the presence of certain traffic engineering sub-TLVs in IS-IS to infer that RSVP signalling is enabled on a given link. A study was conducted with various vendor implementations to determine which traffic engineering sub-TLVs cause an implementation to infer that RSVP signalling is enabled on a link. The results are shown in Figure 1.

TLV/ sub-TLV	Sub-TLV name	Implementation		
		X	Y	Z
22	Extended IS Reachability TLV	N	N	N
22/3	Administrative group (color)	N	Y	Y
22/4	Link Local/Remote ID	N	N	N
22/6	IPv4 Interface Address	N	N	N
22/8	IPv4 Neighbor Address	N	N	N
22/9	Max Link Bandwidth	N	Y	Y
22/10	Max Reservable Link Bandwidth	N	Y	Y
22/11	Unreserved Bandwidth	Y	Y	Y
22/14	Extended Admin Group	N	Y	N
22/18	TE Default Metric	N	N	N
22/20	Link Protection Type	N	Y	Y
22/21	Interface Switching Capability	N	Y	Y
22/22	TE Bandwidth Constraints	N	Y	Y
22/33-39	TE Metric Extensions(RFC7180)	N	N	N
138	SRLG TLV	N	Y	Y

Figure 1: Traffic engineering Sub-TLVs that cause implementation X, Y, or Z to infer that RSVP signalling is enabled on a link

The study indicates that the different implementations use the presence of different sub-TLVs under TLV 22 (or the presence of TLV 138) to infer that RSVP signalling is enabled on a link. It is

possible that other implementations may use other sub-TLVs to infer that RSVP is enabled on a link.

This document defines a standard way to indicate whether or not RSVP, segment routing, or another future protocol is enabled on a link. In this way, implementations will not have to infer whether or not RSVP is enabled based on the presence of different sub-TLVs, but can use the explicit indication. When network operators want to use a non-RSVP traffic engineering application (such as IP/LDP FRR or segment routing), they will be able to advertise traffic engineering sub-TLVs and explicitly indicate what traffic engineering protocols are enabled on a link.

2. Goals

1. The solution should allow the TE protocol enabled on a link to be communicated unambiguously.
2. The solution should decouple the advertisement of which TE protocols are enabled on a link from the advertisement of other TE attributes.
3. The solution should be backward compatible so that nodes that do not understand the new advertisement do not cause issues for existing RSVP deployments.
4. The solution should be extensible for new protocols.
5. The solution should try to limit any increases to the quantity and size of link state advertisements.

2.1. Explicit and unambiguous indication of TE protocol

Communicating unambiguously which TE protocol is enabled on a link is important to be able to share this information with other consumers through other protocols, aside from just the IGP. For example, for a network running both RSVP-TE and SR, it will be useful to communicate which TE protocols are enabled on which links via BGP-LS [RFC7752] to a central controller. Typically, BGP-LS relies on the IGP to distribute IGP topology and traffic engineering information so that only a few BGP-LS sessions with the central controller are needed. In order for a router running a BGP-LS session to a central controller to correctly communicate what TE protocols are enabled on the links in the IGP domain, that information first needs to be communicated unambiguously within the IGP itself. As Figure 1 illustrates, that is currently not the case.

3. Solution

3.1. Traffic-engineering protocol sub-TLV

A new Traffic-engineering protocol sub-TLV is added in the TLV 22 [RFC5305] or TLV 222 to indicate the protocols enabled on the link. The sub-TLV has flags in the value field to indicate the protocol enabled on the link. The length field is variable to allow the flags field to grow for future requirements.

Type : TBD suggested value 40

Length: Variable

Value :

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Flags                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Figure 2: Traffic-Engineering Protocol sub-TLV

Type : TBA (suggested value 40)

Length: variable (in bytes)

Value: The value field consists of bits indicating the protocols enabled on the link. This document defines the two protocol values below.

Value	Protocol Name
0x01	RSVP
0x02	Segment Routing

Figure 3: Flags for the protocols

The RSVP flag is set to one to indicate that RSVP-TE is enabled on a link. The RSVP flag is set to zero to indicate that RSVP-TE is not enabled on a link.

The Segment Routing flag is set to one to indicate that Segment Routing is enabled on a link. The Segment Routing flag is set to zero to indicate that Segment Routing is not enabled on a link.

All undefined flags MUST be set to zero on transmit and ignored on receipt.

An implementation that supports the TE protocol sub-TLV and sends TLV 22 MUST advertise the TE protocol sub-TLV in TLV 22 for that link, even when both the RSVP and SR flags are set to zero. In other words, whenever the TE protocol sub-TLV is supported, it MUST be sent, even if no TE protocols are enabled on the link. This allows a receiving router to determine whether or not the sending router is capable of sending the TE protocol sub-TLV.

A router supporting the TE protocol sub-TLV which receives an advertisement for a link containing TLV 22 with the TE protocol sub-TLV present SHOULD respect the values of the flags in the TE protocol sub-TLV. The receiving router SHOULD only consider links with a given TE protocol enabled for inclusion in a path using that TE protocol. Conversely, links for which the TE protocol sub-TLV is present, but for which the TE protocol flag is not set to one, SHOULD NOT be included in any TE CSPF computations on the receiving router for the protocol in question.

The ability for a receiving router to determine whether or not the sending router is capable of sending the TE protocol sub-TLV is also used for backward compatibility as described in Section 4.

An implementation that supports the TE protocol sub-TLV SHOULD be able to advertise TE sub-TLVs without enabling RSVP-TE signalling on the link.

3.2. Segment Routing flag considerations

The Segment Routing (SR) architecture assumes that the SR topology is congruent with the IGP topology. The path described by a prefix segment is computed using the SPF algorithm applied to the IGP topology, which is the same as the SR topology. Therefore, the presence or absence of the Segment Routing flag MUST NOT be interpreted as modifying the SR topology, which is always congruent with the IGP topology.

It is however useful for a centralized application (or an ingress router) to know whether or not it should expect to be able to forward traffic over a given link using labels distributed via SR. If a link is advertised with the TE protocol sub-TLV and the SR flag set to zero, then a centralized application can assume that traffic sent

with a prefix segment whose path crosses that link is unlikely to be forwarded across that link. With this information, a centralized application can decide to use a different path for that traffic by using a different label stack.

4. Backward compatibility

Routers running older software that do not understand the new Traffic-Engineering protocol sub-TLV will continue to interpret the presence of some sub-TLVs in TLV 22 or the presence of TLV 138 as meaning that RSVP is enabled on a link. A network operator may not want to or be able to upgrade all routers in the domain at the same time. There are two backward compatibility scenarios to consider depending on whether the router that doesn't understand the new TE protocol sub-TLV is an RSVP-TE ingress router or an RSVP-TE transit router.

4.1. Scenario with upgraded RSVP-TE transit router but RSVP-TE ingress router not upgraded

An upgraded RSVP-TE transit router is able to explicitly indicate that RSVP is not enabled on a link by advertising the TE protocol sub-TLV with the RSVP flag set to zero. However, an RSVP-TE ingress router that has not been upgraded to understand the new TE protocol sub-TLV will not understand that RSVP-TE is not enabled on the link, and may include the link on a path computed for RSVP-TE. When the network tries to signal an explicit path LSP using RSVP-TE through that link, it will fail. In order to avoid this scenario, an operator can use the mechanism described below.

For this scenario, the basic idea is to use the existing administrative group link attribute as a means of preventing existing RSVP implementations from using a link. The network operator defines an administrative group to mean that RSVP is not enabled on a link. We call this admin group the RSVP-not-enabled admin group. If the operator needs to advertise a TE sub-TLV (maximum link bandwidth, for example) on a link, but doesn't want to enable RSVP on that link, then the operator also advertises the RSVP-not-enabled admin group on that link. The operator can then use existing mechanisms to exclude links advertising the RSVP-not-enabled admin group from the constrained shortest path first (CSPF) computation used by RSVP. This will prevent RSVP implementations from attempting to signal RSVP-TE LSPs across links that do not have RSVP enabled. Once the entire network domain is upgraded to understand the TE protocol sub-TLV in this draft, the configuration involving the RSVP-not-enabled admin group is no longer needed for this network.

4.2. Scenario with upgraded RSVP-TE ingress router but RSVP-TE transit router not upgraded

The other scenario to consider is when the RSVP-TE ingress router has been upgraded to understand the TE protocol sub-TLV, but the RSVP-TE transit router has not. In this case, the transit router has not been upgraded, so it is not yet capable of sending the TE protocol sub-TLV. If the transit router has RSVP-TE enabled on a link, we would like for the RSVP-TE ingress router to still be able to use the link for RSVP-TE paths. While it is possible to describe a solution for this scenario that makes use of administrative groups, we describe a simpler solution below.

The solution for this scenario relies on the following observation. If the RSVP-TE ingress router can understand that the transit router is not capable of sending the TE protocol sub-TLV, then it can continue inferring whether or not RSVP-TE is enabled on the transit router links based on the presence of TE sub-TLVs, just as it does today.

To accomplish this, we require an upgraded router to send the TE protocol sub-TLV if it sends TLV 22, even when both the RSVP and SR flags are set to zero. In other words, whenever the TE protocol sub-TLV is supported, it MUST be sent, even if no TE protocols are enabled on the link. see Section 3. This allows the receiving router to interpret the absence of the TE-protocol sub-TLV together with presence of TLV 22 to mean that the sending router has not been upgraded. This allows the upgraded RSVP-TE ingress router to distinguish between transit routers that have been upgraded and those that haven't. When the transit router has been upgraded, then the RSVP-TE ingress router uses the information in the TE protocol sub-TLV. When the transit router has not been upgraded, then RSVP-TE ingress router continues to infer whether or not RSVP-TE is enabled on the transit router links based on the presence of TE sub-TLVs, just as it does today. The solution for this scenario requires no configuration on the part of network operators.

4.3. Need for a long term solution

The use of the administrative group link attribute to prevent an RSVP-TE ingress router from computing a path using a given link is an effective short term workaround to allow networks to incrementally upgrade the routers to software that understands the new TE-protocol sub-TLV. One might also consider a long term solution based solely on the use of operator-defined administrative groups to communicate the TE protocol enabled on a link. However, we do not consider this workaround to be an effective long term solution because it relies on operator configuration that would have to be maintained in the long

term. As discussed in Section 2, continuing to have to infer which TE protocol is enabled on a link also limits our ability to communicate this information unambiguously in an interoperable manner for use by other applications such as central controllers.

5. Security Considerations

This document does not introduce any further security issues other than those discussed in [RFC1195] and [RFC5305].

6. IANA Considerations

This specification updates one IS-IS registry:

The extended IS reachability TLV Registry

i) Traffic-engineering Protocol sub-tlv = Suggested value 40

7. Acknowledgements

The authors thank Alia Atlas, Les Ginsberg, and Peter Psenak for helpful discussions on the topic of this draft.

8. References

8.1. Normative References

- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S.,
and R. Shakir, "Segment Routing Architecture", draft-ietf-
spring-segment-routing-09 (work in progress), July 2016.
- [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>>.
- [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>>.
- [RFC7810] Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and
Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions",
RFC 7810, DOI 10.17487/RFC7810, May 2016,
<<https://www.rfc-editor.org/info/rfc7810>>.

8.2. Informative 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, <<https://www.rfc-editor.org/info/rfc1195>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [RFC7916] Litkowski, S., Ed., Decraene, B., Filsfils, C., Raza, K., Horneffer, M., and P. Sarkar, "Operational Management of Loop-Free Alternates", RFC 7916, DOI 10.17487/RFC7916, July 2016, <<https://www.rfc-editor.org/info/rfc7916>>.

Authors' Addresses

Shraddha Hegde
Juniper Networks
Embassy Business Park
Bangalore, KA 560093
India

Email: shraddha@juniper.net

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: cbowers@juniper.net

Paul Mattes
Microsoft
One Microsoft Way
Redmond, WA 98052
US

Email: pamattes@microsoft.com

Mohan Nanduri
Microsoft
One Microsoft Way
Redmond, WA 98052
US

Email: mnanduri@microsoft.com

Spencer Giacalone
Microsoft
One Microsoft Way
Redmond, WA 98052
US

Email: Spencer.Giacalone@microsoft.com

Imtiyaz Mohammad
Arista Networks
Global Tech Park
Bangalore, KA 560103
India

Email: imtiyaz@arista.com

Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: June 6, 2019

N. Shen
Cisco Systems
S. Amante
Apple, Inc.
M. Abrahamsson
T-Systems Nordic
December 3, 2018

IS-IS Routing with Reverse Metric
draft-ietf-isis-reverse-metric-17

Abstract

This document describes a mechanism to allow IS-IS routing to quickly and accurately shift traffic away from either a point-to-point or multi-access LAN interface during network maintenance or other operational events. This is accomplished by signaling adjacent IS-IS neighbors with a higher reverse metric, i.e., the metric towards the signaling IS-IS router.

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 June 6, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Node and Link Isolation	3
1.2. Distributed Forwarding Planes	3
1.3. Spine-Leaf Applications	3
1.4. LDP IGP Synchronization	3
1.5. IS-IS Reverse Metric	4
1.6. Specification of Requirements	4
2. IS-IS Reverse Metric TLV	4
3. Elements of Procedure	7
3.1. Processing Changes to Default Metric	7
3.2. Multi-Topology IS-IS Support on Point-to-point links . .	7
3.3. Multi-Access LAN Procedures	7
3.4. LDP/IGP Synchronization on LANs	9
3.5. Operational Guidelines	9
4. Security Considerations	10
5. IANA Considerations	10
6. Acknowledgments	11
7. References	11
7.1. Normative References	11
7.2. Informative References	12
Appendix A. Node Isolation Challenges	12
Appendix B. Link Isolation Challenges	13
Appendix C. Contributors' Addresses	14
Authors' Addresses	14

1. Introduction

The IS-IS [ISO10589] routing protocol has been widely used in Internet Service Provider IP/MPLS networks. Operational experience with the protocol, combined with ever increasing requirements for lossless operations have demonstrated some operational issues. This document describes the issues and a mechanism for mitigating them.

This document defines the IS-IS "Reverse Metric" mechanism that allows an IS-IS node to send a "Reverse Metric" TLV through the IS-IS Hello (IIH) PDU to the neighbor or pseudo-node to adjust the routing metric on the inbound direction.

1.1. Node and Link Isolation

IS-IS routing mechanism has the overload-bit, which can be used by operators to perform disruptive maintenance on the router. But in many operational maintenance cases, it is not necessary to divert all the traffic away from this node. It is necessary to avoid only a single link during the maintenance. More detailed descriptions of the challenges can be found in Appendix A and Appendix B of this document.

1.2. Distributed Forwarding Planes

In a distributed forwarding platform, different forwarding line-cards may have interfaces and IS-IS connections to neighbor routers. If one of the line-card's software resets, it may take some time for the forwarding entries to be fully populated on the line-card, in particular if the router is a PE (Provider Edge) router in ISP's MPLS VPN. An IS-IS adjacency may be established with a neighbor router long before the entire BGP VPN prefixes are downloaded to the forwarding table. It is important to signal to the adjacent IS-IS routers to raise metric values and not to use the corresponding IS-IS adjacency inbound to this router if possible. Temporarily signaling the 'Reverse Metric' over this link to discourage the traffic via the corresponding line-card will help to reduce the traffic loss in the network. In the meantime, the remote PE routers will select a different set of PE routers for the BGP best path calculation or use a different link towards the same PE router on which a line-card is resetting.

1.3. Spine-Leaf Applications

In the IS-IS Spine-Leaf extension [I-D.shen-isis-spine-leaf-ext], the leaf nodes will perform equal-cost or unequal-cost load sharing towards all the spine nodes. In certain operational cases, for instance, when one of the backbone links on a spine node is congested, a spine node can push a higher metric towards the connected leaf nodes to reduce the transit traffic through the corresponding spine node or link.

1.4. LDP IGP Synchronization

In the [RFC5443], a mechanism is described to achieve LDP IGP synchronization by using the maximum link metric value on the interface. But in the case of a new IS-IS node joining the broadcast network (LAN), it is not optimal to change all the nodes on the LAN to the maximum link metric value, as described in [RFC6138]. In this case, the Reverse Metric can be used to discourage both outbound and

inbound traffic without affecting the traffic of other IS-IS nodes on the LAN.

1.5. IS-IS Reverse Metric

This document uses the routing protocol itself as the transport mechanism to allow one IS-IS router to advertise a "reverse metric" in an IS-IS Hello (IIH) PDU to an adjacent node on a point-to-point or multi-access LAN link. This would allow the provisioning to be performed only on a single node, setting a "reverse metric" on a link and have traffic bidirectionally shift away from that link gracefully to alternate, viable paths.

This Reverse Metric mechanism is used for both point-to-point and multi-access LAN links. Unlike the point-to-point links, the IS-IS protocol currently does not have a way to influence the traffic towards a particular node on LAN links. This mechanism provides IS-IS routing the capability of altering traffic in both directions on either a point-to-point link or a multi-access link of an IS-IS node.

The metric value in the "reverse metric" TLV and the Traffic Engineering metric in the sub-TLV being advertised is an offset or relative metric to be added to the existing local link and Traffic Engineering metric values of the receiver, the accumulated metric value is bounded as described in Section 2.

1.6. Specification of Requirements

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

2. IS-IS Reverse Metric TLV

The Reverse Metric TLV is a new TLV to be used inside an IS-IS Hello PDU. This TLV is used to support the IS-IS Reverse Metric mechanism that allows a "reverse metric" to be sent to the IS-IS neighbor.

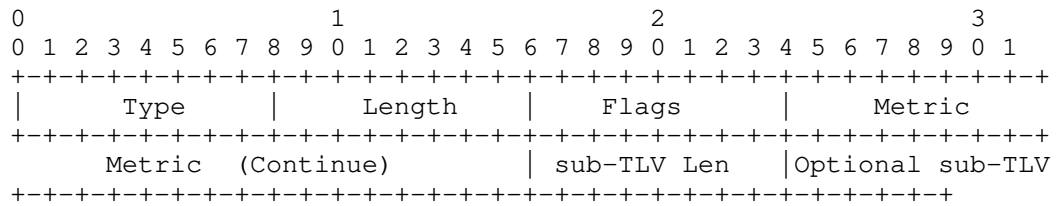


Figure 1: Reverse Metric TLV

The Value part of the Reverse Metric TLV is composed of a 3 octet field containing an IS-IS Metric Value, a 1 octet field of Flags, and a 1 octet Reverse Metric sub-TLV length field representing the length of a variable number of sub-TLVs. If the "sub-TLV len" is non-zero, then the Value field MUST also contain one or more sub-TLVs.

The Reverse Metric TLV MAY be present in any IS-IS Hello PDU. A sender MUST only transmit a single Reverse Metric TLV in a IS-IS Hello PDU. If a received IS-IS Hello PDU contains more than one Reverse Metric TLV, an implementation MUST ignore all the Reverse Metric TLVs.

TYPE: 16
 LENGTH: variable (5 - 255 octets)
 VALUE:

Flags (1 octet)
 Metric (3 octets)
 sub-TLV length (1 octet)
 sub-TLV data (0 - 250 octets)

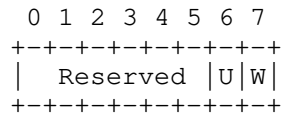


Figure 2: Flags

The Metric field contains a 24-bit unsigned integer. This value is a metric offset that a neighbor SHOULD add to the existing, configured Default Metric for the IS-IS link [ISO10589]. Refer to "Elements of Procedure", in Section 3 for details on how an IS-IS router should process the Metric field in a Reverse Metric TLV.

The Metric field, in the Reverse Metric TLV, is a "reverse offset metric" that will either be in the range of 0 - 63 when a "narrow" IS-IS metric is used (IS Neighbors TLV, Pseudonode LSP) [RFC1195] or in the range of 0 - ($2^{24} - 2$) when a "wide" Traffic Engineering

metric value is used, (Extended IS Reachability TLV) [RFC5305] [RFC5817]. As described below, when the U bit is set, the accumulated value of the wide metric is in the range of 0 - ($2^{24} - 1$), with ($2^{24} - 1$) metric as non-reachable in IS-IS routing. The IS-IS metric value of ($2^{24} - 2$) serves as the link of last resort.

There are currently only two Flag bits defined.

W bit (0x01): The "Whole LAN" bit is only used in the context of multi-access LANs. When a Reverse Metric TLV is transmitted from a node to the Designated Intermediate System (DIS), if the "Whole LAN" bit is set (1), then a DIS SHOULD add the received Metric value in the Reverse Metric TLV to each node's existing Default Metric in the Pseudonode LSP. If the "Whole LAN" bit is not set (0), then a DIS SHOULD add the received Metric value in the Reverse Metric TLV to the existing "default metric" in the Pseudonode LSP for the single node from whom the Reverse Metric TLV was received. Please refer to "Multi-Access LAN Procedures", in Section 3.3, for additional details. The W bit MUST be clear when a Reverse Metric TLV is transmitted in an IIH PDU on a point-to-point link, and MUST be ignored when received on a point-to-point link.

U bit (0x02): The "Unreachable" bit specifies that the metric calculated by addition of the reverse metric to the "default metric" is limited to the maximum value of ($2^{24}-1$). This "U" bit applies to both the default metric in the Extended IS Reachability TLV and the Traffic Engineering Default Metric sub-TLV of the link. This is only relevant to the IS-IS "wide" metric mode.

The Reserved bits of Flags field MUST be set to zero and MUST be ignored when received.

The Reverse Metric TLV MAY include sub-TLVs when an IS-IS router wishes to signal additional information to its neighbor. In this document, the Reverse Metric Traffic Engineering Metric sub-TLV, with Type 18, is defined. This Traffic Engineering Metric contains a 24-bit unsigned integer. This sub-TLV is optional, if it appears more than once, then the entire Reverse Metric TLV MUST be ignored. Upon receiving this Traffic Engineering METRIC sub-TLV in a Reverse Metric TLV, a node SHOULD add the received Traffic Engineering Metric offset value to its existing, configured Traffic Engineering Default Metric within its Extended IS Reachability TLV. The use of other sub-TLVs is outside the scope of this document. The "sub-TLV Len" value MUST be set to zero when an IS-IS router does not have Traffic Engineering sub-TLVs that it wishes to send to its IS-IS neighbor.

3. Elements of Procedure

3.1. Processing Changes to Default Metric

It is important to use the same IS-IS metric type on both ends of the link and in the entire IS-IS area or level. On the receiving side of the 'reverse-metric' TLV, the accumulated value of configured metric and the reverse-metric needs to be limited to 63 in "narrow" metric mode and to $(2^{24} - 2)$ in "wide" metric mode. This applies to both the Default Metric of Extended IS Reachability TLV and the Traffic Engineering Default Metric sub-TLV in LSP or Pseudonode LSP for the "wide" metric mode case. If the "U" bit is present in the flags, the accumulated metric value is to be limited to $(2^{24} - 1)$ for both the normal link metric and Traffic Engineering metric in IS-IS "wide" metric mode.

If an IS-IS router is configured to originate a Traffic Engineering Default Metric sub-TLV for a link, but receives a Reverse Metric TLV from its neighbor that does not contain a Traffic Engineering Default Metric sub-TLV, then the IS-IS router MUST NOT change the value of its Traffic Engineering Default Metric sub-TLV for that link.

3.2. Multi-Topology IS-IS Support on Point-to-point links

The Reverse Metric TLV is applicable to Multi-Topology IS-IS (M-ISIS) [RFC5120]. On point-to-point links, if an IS-IS router is configured for M-ISIS, it MUST send only a single Reverse Metric TLV in IIH PDUs toward its neighbor(s) on the designated link. When an M-ISIS router receives a Reverse Metric TLV, it MUST add the received Metric value to its Default Metric of the link in all Extended IS Reachability TLVs for all topologies. If an M-ISIS router receives a Reverse Metric TLV with a Traffic Engineering Default Metric sub-TLV, then the M-ISIS router MUST add the received Traffic Engineering Default Metric value to each of its Default Metric sub-TLVs in all of its MT Intermediate Systems TLVs. If an M-ISIS router is configured to advertise Traffic Engineering Default Metric sub-TLVs for one or more topologies, but does not receive a Traffic Engineering Default Metric sub-TLV in a Reverse Metric TLV, then the M-ISIS router MUST NOT change the value in each of the Traffic Engineering Default Metric sub-TLVs for all topologies.

3.3. Multi-Access LAN Procedures

On a Multi-Access LAN, only the DIS SHOULD act upon information contained in a received Reverse Metric TLV. All non-DIS nodes MUST silently ignore a received Reverse Metric TLV. The decision process of the routers on the LAN MUST follow the procedure in section

7.2.8.2 of [ISO10589], and use the "Two-way connectivity check" during the topology and route calculation.

The Reverse Metric Traffic Engineering sub-TLV also applies to the DIS. If a DIS is configured to apply Traffic Engineering over a link and it receives Traffic Engineering Metric sub-TLV in a Reverse Metric TLV, it should update the Traffic Engineering Default Metric sub-TLV value of the corresponding Extended IS Reachability TLV or insert a new one if not present.

In the case of multi-access LANs, the "W" Flags bit is used to signal from a non-DIS to the DIS whether to change the metric and, optionally, Traffic Engineering parameters for all nodes in the Pseudonode LSP or solely the node on the LAN originating the Reverse Metric TLV.

A non-DIS node, e.g., Router B, attached to a multi-access LAN will send the DIS a Reverse Metric TLV with the W bit clear when Router B wishes the DIS to add the Metric value to the Default Metric contained in the Pseudonode LSP specific to just Router B. Other non-DIS nodes, e.g., Routers C and D, may simultaneously send a Reverse Metric TLV with the W bit clear to request the DIS to add their own Metric value to their Default Metric contained in the Pseudonode LSP.

As long as at least one IS-IS node on the LAN sending the signal to DIS with the W bit set, the DIS would add the metric value in the Reverse Metric TLV to all neighbor adjacencies in the Pseudonode LSP, regardless if some of the nodes on the LAN advertise the Reverse Metric TLV without the W bit set. The DIS MUST use the reverse metric of the highest source MAC address Non-DIS advertising the Reverse Metric TLV with the W bit set.

Local provisioning on the DIS to adjust the Default Metric(s) is another way to insert Reverse Metric in the Pseudonode LSP towards an IS-IS node on a LAN. In the case where Reverse Metric TLV is also used in the IS-IS Hello PDU of the node, the local provisioning MUST take precedence over received Reverse Metric TLVs. For instance, local policy on the DIS may be provisioned to ignore the W bit signaling on a LAN.

Multi-Topology IS-IS [RFC5120] specifies there is no change to construction of the Pseudonode LSP, regardless of the Multi-Topology capabilities of a multi-access LAN. If any MT capable node on the LAN advertises the Reverse Metric TLV to the DIS, the DIS should update, as appropriate, the Default Metric contained in the Pseudonode LSP. If the DIS updates the Default Metric in and floods

a new Pseudonode LSP, those default metric values will be applied to all topologies during Multi-Topology SPF calculations.

3.4. LDP/IGP Synchronization on LANs

As described in [RFC6138] when a new IS-IS node joins a broadcast network, it is unnecessary and sometimes even harmful for all IS-IS nodes on the LAN to advertise maximum link metric. [RFC6138] proposes a solution to have the new node not advertise its adjacency towards the pseudo-node when it is not in a "cut-edge" position.

With the introduction of Reverse Metric in this document, a simpler alternative solution to the above mentioned problem can be used. The Reverse Metric allows the new node on the LAN to advertise its inbound metric value to be the maximum and this puts the link of this new node in the last resort position without impacting the other IS-IS nodes on the same LAN.

Specifically, when IS-IS adjacencies are being established by the new node on the LAN, besides setting the maximum link metric value ($2^{24} - 2$) on the interface of the LAN for LDP IGP synchronization as described in [RFC5443], it SHOULD advertise the maximum metric offset value in the Reverse Metric TLV in its IIH PDU sent on the LAN. It SHOULD continue this advertisement until it completes all the LDP label binding exchanges with all the neighbors over this LAN, either by receiving the LDP End-of-LIB [RFC5919] for all the sessions or by exceeding the provisioned timeout value for the node LDP/IGP synchronization.

3.5. Operational Guidelines

For the use case in Section 1.1, a router SHOULD limit the period of advertising a Reverse Metric TLV towards a neighbor only for the duration of network maintenance window.

The use of Reverse Metric does not alter IS-IS metric parameters stored in a router's persistent provisioning database.

If routers that receive a Reverse Metric TLV sends a syslog message or SNMP trap, this will assist in rapidly identifying the node in the network that is advertising an IS-IS metric or Traffic Engineering parameters different from that which is configured locally on the device.

When the link Traffic Engineering metric is raised to ($2^{24} - 1$) [RFC5817], either due to the reverse-metric mechanism or by explicit user configuration, this SHOULD immediately trigger the CSPF (Constrained Shortest Path First) re-calculation to move the Traffic

Engineering traffic away from that link. It is RECOMMENDED also that the CSPF does the immediate CSPF re-calculation when the Traffic Engineering metric is raised to $(2^{24} - 2)$ to be the last resort link.

It is advisable that implementations provide a configuration capability to disable any IS-IS metric changes by Reverse Metric mechanism through neighbor's Hello PDUs.

If an implementation enables this mechanism by default, it is RECOMMENDED that it be disabled by the operators when not explicitly using it.

4. Security Considerations

Security concerns for IS-IS are addressed in [ISO10589], [RFC5304], [RFC5310], and with various deployment and operational security considerations in [RFC7645]. The enhancement in this document makes it possible for one IS-IS router to manipulate the IS-IS Default Metric and, optionally, Traffic Engineering parameters of adjacent IS-IS neighbors on point-to-point or LAN interfaces. Although IS-IS routers within a single Autonomous System nearly always are under the control of a single administrative authority, it is highly recommended that operators configure authentication of IS-IS PDUs to mitigate use of the Reverse Metric TLV as a potential attack vector.

5. IANA Considerations

IANA has allocated IS-IS TLV Codepoints of 16 for the Reverse Metric TLV. This new TLV has the following attributes: IIH = y, LSP = n, SNP = n, Purge = n.

This document also introduces a new registry for sub-TLVs of the Reverse Metric TLV. The registration policy is Expert Review as defined in [RFC8126]. This registry is part of the "IS-IS TLV Codepoints" registry. The name of the registry is "Sub-TLVs for Reverse Metric TLV". The defined values are:

0:	Reserved
1-17:	Unassigned
18:	Traffic Engineering Metric sub-TLV, as specified in this document (Section 2)
19-255:	Unassigned

6. Acknowledgments

The authors would like to thank Mike Shand, Dave Katz, Guan Deng, Ilya Varlashkin, Jay Chen, Les Ginsberg, Peter Ashwood-Smith, Uma Chunduri, Alexander Okonnikov, Jonathan Harrison, Dave Ward, Himanshu Shah, Wes George, Danny McPherson, Ed Crabbe, Russ White, Robert Raszuk, Tom Petch, Stewart Bryant and Acee Lindem for their comments and contributions.

This document was produced using Marshall Rose's xml2rfc tool.

7. References

7.1. Normative References

- [ISO10589] ISO, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002.
- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <<https://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, <<https://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, <<https://www.rfc-editor.org/info/rfc5120>>.
- [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>>.
- [RFC5443] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, DOI 10.17487/RFC5443, March 2009, <<https://www.rfc-editor.org/info/rfc5443>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.

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

7.2. Informative References

- [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-07 (work in progress), October 2018.
- [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>>.
- [RFC5817] Ali, Z., Vasseur, JP., Zamfir, A., and J. Newton, "Graceful Shutdown in MPLS and Generalized MPLS Traffic Engineering Networks", RFC 5817, DOI 10.17487/RFC5817, April 2010, <<https://www.rfc-editor.org/info/rfc5817>>.
- [RFC5919] Asati, R., Mohapatra, P., Chen, E., and B. Thomas, "Signaling LDP Label Advertisement Completion", RFC 5919, DOI 10.17487/RFC5919, August 2010, <<https://www.rfc-editor.org/info/rfc5919>>.
- [RFC6138] Kini, S., Ed. and W. Lu, Ed., "LDP IGP Synchronization for Broadcast Networks", RFC 6138, DOI 10.17487/RFC6138, February 2011, <<https://www.rfc-editor.org/info/rfc6138>>.
- [RFC7645] Chunduri, U., Tian, A., and W. Lu, "The Keying and Authentication for Routing Protocol (KARP) IS-IS Security Analysis", RFC 7645, DOI 10.17487/RFC7645, September 2015, <<https://www.rfc-editor.org/info/rfc7645>>.

Appendix A. Node Isolation Challenges

On rare occasions, it is necessary for an operator to perform disruptive network maintenance on an entire IS-IS router node, i.e., major software upgrades, power/cooling augments, etc. In these cases, an operator will set the IS-IS Overload Bit (OL-bit) within the Link State Protocol Data Units (LSPs) of the IS-IS router about to undergo maintenance. The IS-IS router immediately floods its updated LSPs to all IS-IS routers in the IS-IS domain. Upon receipt

of the updated LSPs, all IS-IS routers recalculate their Shortest Path First (SPF) tree excluding IS-IS routers whose LSPs have the OL-bit set. This effectively removes the IS-IS router about to undergo maintenance from the topology, thus preventing it from receiving any transit traffic during the maintenance period.

After the maintenance activity has completed, the operator resets the IS-IS Overload Bit within the LSPs of the original IS-IS router causing it to flood updated IS-IS LSPs throughout the IS-IS domain. All IS-IS routers recalculate their SPF tree and now include the original IS-IS router in their topology calculations, allowing it to be used for transit traffic again.

Isolating an entire IS-IS router from the topology can be especially disruptive due to the displacement of a large volume of traffic through an entire IS-IS router to other, sub-optimal paths, (e.g., those with significantly larger delay). Thus, in the majority of network maintenance scenarios, where only a single link or LAN needs to be augmented to increase its physical capacity or is experiencing an intermittent failure, it is much more common and desirable to gracefully remove just the targeted link or LAN from service, temporarily, so that the least amount of user-data traffic is affected during the link-specific network maintenance.

Appendix B. Link Isolation Challenges

Before network maintenance events are performed on individual physical links or LANs, operators substantially increase the IS-IS metric simultaneously on both devices attached to the same link or LAN. In doing so, the devices generate new Link State Protocol Data Units (LSPs) that are flooded throughout the network and cause all routers to gradually shift traffic onto alternate paths with very little or no disruption to in-flight communications by applications or end-users. When performed successfully, this allows the operator to confidently perform disruptive augmentation, fault diagnosis or repairs on a link without disturbing ongoing communications in the network.

There are a number of challenges with the above solution. First, it is quite common to have routers with several hundred interfaces and individual interfaces that are from several hundred Gigabits/second to Terabits/second of traffic. Thus, it is imperative that operators accurately identify the same point-to-point link on two, separate devices in order to increase (and, afterward, decrease) the IS-IS metric appropriately. Second, the aforementioned solution is very time consuming and even more error-prone to perform when it's necessary to temporarily remove a multi-access LAN from the network topology. Specifically, the operator needs to configure ALL devices

that have interfaces attached to the multi-access LAN with an appropriately high IS-IS metric, (and then decrease the IS-IS metric to its original value afterward). Finally, with respect to multi-access LANs, there is currently no method to bidirectionally isolate only a single node's interface on the LAN when performing more fine-grained diagnosis and repairs to the multi-access LAN.

In theory, use of a Network Management System (NMS) could improve the accuracy of identifying the appropriate subset of routers attached to either a point-to-point link or a multi-access LAN as well as signaling from the NMS to those devices, using a network management protocol to adjust the IS-IS metrics on the pertinent set of interfaces. The reality is that NMSs are, to a very large extent, not used within Service Provider's networks for a variety of reasons. In particular, NMSs do not interoperate very well across different vendors or even separate platform families within the same vendor.

Appendix C. Contributors' Addresses

Tony Li

Email: tony.li@tony.li

Authors' Addresses

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

Email: naiming@cisco.com

Shane Amante
Apple, Inc.
1 Infinite Loop
Cupertino, CA 95014
USA

Email: samante@apple.com

Mikael Abrahamsson
T-Systems Nordic
Kistagangen 26
Stockholm
SE

Email: Mikael.Abrahamsson@t-systems.se

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).

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 19, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
2. Motivations	3
3. Spine-Leaf (SL) Extension	4
3.1. Topology Examples	4
3.2. Applicability Statement	5
3.3. Spine-Leaf TLV	6
3.3.1. Spine-Leaf Sub-TLVs	7
3.3.1.1. Leaf-Set Sub-TLV	7
3.3.1.2. Info-Req Sub-TLV	8
3.3.2. Advertising IPv4/IPv6 Reachability	8
3.3.3. Advertising Connection to RF-Leaf Node	8
3.4. Mechanism	8
3.4.1. Pure CLOS Topology	10
3.5. Implementation and Operation	11
3.5.1. CSNP PDU	11
3.5.2. Overload Bit	11
3.5.3. Spine Node Hostname	11
3.5.4. IS-IS Reverse Metric	11
3.5.5. Spine-Leaf Traffic Engineering	12
3.5.6. Other End-to-End Services	12
3.5.7. Address Family and Topology	12
3.5.8. Migration	13
4. IANA Considerations	13
5. Security Considerations	14
6. Acknowledgments	14
7. Document Change Log	14
7.1. Changes to draft-shen-isis-spine-leaf-ext-05.txt	14
7.2. Changes to draft-shen-isis-spine-leaf-ext-04.txt	14
7.3. Changes to draft-shen-isis-spine-leaf-ext-03.txt	14
7.4. Changes to draft-shen-isis-spine-leaf-ext-02.txt	14
7.5. Changes to draft-shen-isis-spine-leaf-ext-01.txt	15
7.6. Changes to draft-shen-isis-spine-leaf-ext-00.txt	15
8. References	15
8.1. Normative References	15
8.2. Informative References	16
Authors' Addresses	17

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

- o 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
- o 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.

- o 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.
- o 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.
- o 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

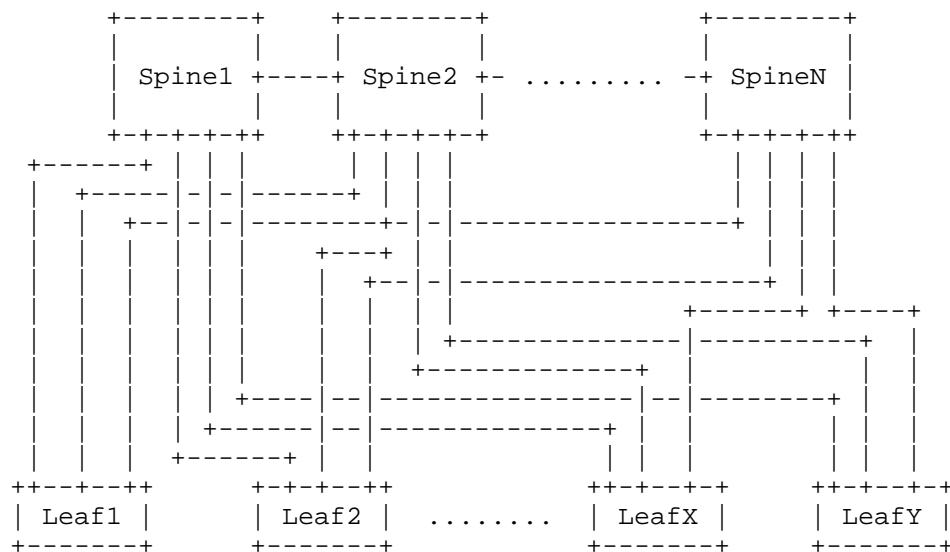


Figure 1: A Spine-Leaf Topology

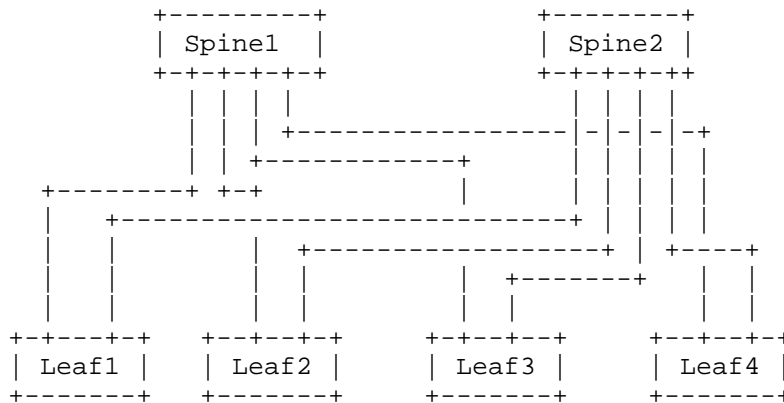


Figure 2: A CLOS Topology

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.

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-----+-----+-----+-----+-----+-----+-----+-----+
    |          Type          |      Length      |          SL Flag          |
    +-----+-----+-----+-----+-----+-----+-----+-----+
    |          .. Optional Sub-TLVs          |
    +-----+-----+-----+-----+-----+-----+

```

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

```

    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
    +-----+-----+-----+-----+-----+-----+
    | Tier |      Reserved      | T | R | L |
    +-----+-----+-----+-----+-----+-----+

```

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

- o Submitted January 2018.
- o Just a refresh.

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

- o Submitted June 2017.
- o 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

- o Submitted March 2017.
- o Added the Spine-Leaf sub-TLVs to handle the case of data center pure CLOS topology and mechanism.
- o Added the Spine-Leaf TLV and sub-TLVs can be optionally inserted in either IIH PDU or CS-LSP PDU.
- o 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

- o Submitted October 2016.
- o 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

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

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

- o 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
routing information exchange protocol for use in
conjunction with the protocol for providing the
connectionless-mode Network Service (ISO 8473), ISO/IEC
10589:2002, Second Edition.", Nov 2002.

[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

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 9, 2019

R. White, Ed.
S. Zandi, Ed.
LinkedIn
November 5, 2018

IS-IS Support for Openfabric
draft-white-openfabric-07

Abstract

Spine and leaf topologies are widely used in hyperscale and cloud scale networks. In most of these networks, configuration is automated, but difficult, and topology information is extracted through broad based connections. Policy is often integrated into the control plane, as well, making configuration, management, and troubleshooting difficult. Openfabric is an adaptation of an existing, widely deployed link state protocol, Intermediate System to Intermediate System (IS-IS) that is designed to:

- o Provide a full view of the topology from a single point in the network to simplify operations
- o Minimize configuration of each Intermediate System (IS) (also called a router or switch) in the network
- o Optimize the operation of IS-IS within a spine and leaf fabric to enable scaling

This document begins with an overview of openfabric, including a description of what may be removed from IS-IS to enable scaling. The document then describes an optimized adjacency formation process; an optimized flooding scheme; some thoughts on the operation of openfabric, metrics, and aggregation; and finally a description of the changes to the IS-IS protocol required for openfabric.

Status of This Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 9, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Goals	3
1.2. Contributors	3
1.3. Simplification	3
1.4. Additions and Requirements	4
1.5. Sample Network	4
2. Modified Adjacency Formation	6
2.1. Level 2 Adjacencies Only	6
2.2. Point-to-point Adjacencies	6
2.3. Three Way Handshake Support	7
2.4. Adjacency Formation Optimization	7
3. Advertisement of Reachability Information	8
4. Determining and Advertising Location on the Fabric	9
5. Flooding Optimization	10
5.1. Flooding Failures	11
6. Other Optimizations	12
6.1. Transit Link Reachability	12
6.2. Transiting T0 Intermediate Systems	12
7. Openfabric and Route Aggregation	13
8. Security Considerations	13
9. References	13
9.1. Normative References	13
9.2. Informative References	15
Appendix A. Flooding Optimization Operation	17
Appendix B. Fabric Location Calculation	19
Authors' Addresses	20

1. Introduction

1.1. Goals

Spine and leaf fabrics are often used in large scale data centers; in this application, they are commonly called a fabric because of their regular structure and predictable forwarding and convergence properties. This document describes modifications to the IS-IS protocol to enable it to run efficiently on a large scale spine and leaf fabric, openfabric. The goals of this control plane are:

- o Provide a full view of the topology from a single point in the network to simplify operations
- o Minimize configuration of each IS in the network
- o Optimize the operation of IS-IS within a spine and leaf fabric to enable scaling

1.2. Contributors

The following people have contributed to this draft: Nikos Triantafyllis (reflected flooding optimization), Ivan Pepelnjak (fabric locality calculation modifications), Christian Franke (fabric locality calculation modification), Hannes Gredler (do not reflood optimizations), Les Ginsberg (capabilities encoding, circuit local reflooding), Naiming Shen (capabilities encoding, circuit local reflooding), Uma Chunduri (failure mode suggestions, flooding), Nick Russo, and Rodny Molina.

See [RFC5449], [RFC5614], and [RFC7182] for similar solutions in the Mobile Ad Hoc Networking (MANET) solution space.

1.3. Simplification

In building any scalable system, it is often best to begin by removing what is not needed. In this spirit, openfabric implementations MAY remove the following from IS-IS:

- o External metrics. There is no need for external metrics in large scale spine and leaf fabrics; it is assumed that metrics will be properly configured by the operator to account for the correct order of route preference at any route redistribution point.
- o Tags and traffic engineering processing. Openfabric is only designed to provide topology and reachability information. It is not designed to provide for traffic engineering, route preference through tags, or other policy mechanisms. It is assumed that all

routing policy will be provided through an overlay system which communicates directly with each IS in the fabric, such as PCEP [RFC5440] or I2RS [RFC7921]. Traffic engineering is assumed to be provided through Segment Routing (SR) [I-D.ietf-spring-segment-routing].

1.4. Additions and Requirements

To create a scalable link state fabric, openfabric includes the following:

- o A slightly modified adjacency formation process.
- o Mechanisms for determining which tier within a spine and leaf fabric in which the IS is located.
- o A mechanism that reduces flooding to the minimum possible, while still ensuring complete database synchronization among the intermediate systems within the fabric.

Three general requirements are placed here; more specific requirements are considered in the following sections. Openfabric implementations:

- o MUST support [RFC5301] and enable hostname advertisement by default if a hostname is configured on the intermediate system.
- o SHOULD support [RFC6232], purge originator identification for IS-IS.
- o MUST NOT be mixed with standard IS-IS implementations in operational deployments. Openfabric and standard IS-IS implementations SHOULD be treated as two separate protocols.

1.5. Sample Network

The following spine and leaf fabric will be used to describe these modifications.

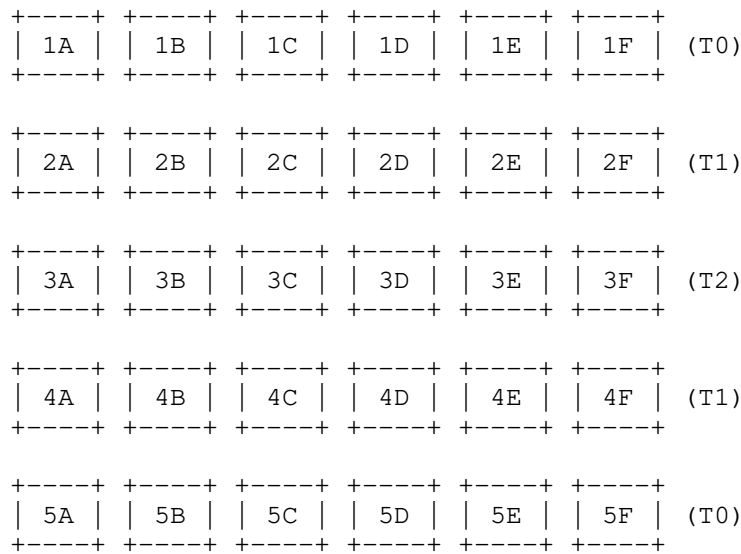


Figure 1

To reduce confusion (spine and leaf fabrics are difficult to draw in plain text art), this diagram does not contain the connections between devices. The reader should assume that each device in a given layer is connected to every device in the layer above it. For instance:

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

The tiers or stages of the fabric are also marked for easier reference. T0 is assumed to be connected to application servers, or rather they are Top of Rack (ToR) intermediate systems. The remaining tiers, T1 and T2, are connected only to the fabric itself. Note there are no "cross links," or "east west" links in the illustrated fabric. The fabric locality detection mechanism described here will not work if there are cross links running east/

west through the fabric. Locality detection may be possible in such a fabric; this is an area for further study.

2. Modified Adjacency Formation

Because Openfabric operates in a tightly controlled data center environment, various modifications can be made to the IS-IS neighbor formation process to increase efficiency and simplify the protocol. Specifically, Openfabric implementations SHOULD support [RFC3719], section 4, hello padding for IS-IS. Variable hello padding SHOULD NOT be used, as data center fabrics are built using high speed links on which padded hellos will have little performance impact. Further modifications to the neighbor formation process are considered in the following sections.

2.1. Level 2 Adjacencies Only

Openfabric is designed to work in a single flooding domain over a single data center fabric at the scale of thousands of routers with hundreds of thousands of routes (so a moderate scale in router and route count terms). Because of the way Openfabric optimizes operation in this environment, it is not necessary nor desirable to build multiple flooding domains. For instance, the flooding optimizations described later in this document require a full view of the topology, as does any proposed overlay to inject policy into the forwarding plane. In light of this, the following changes SHOULD BE to IS-IS implementations to support Openfabric:

- o IIH PDU 17 (level 2 point-to-point circuit hello) should be the only IIH PDU type transmitted (see section 9.7 of ISO 10589)
- o In IIH PDU 17 (level 2 point-to-point circuit hello), the Circuit Type field should be set to 2 (see section 9.7 of ISO 10589)
- o Support for IIH PDU 15 (level 1 broadcast hello) should be removed (see section 9.5 of ISO 10589)
- o Support for IIH PDU 16 (level 2 broadcast hello) should be removed (see section 9.6 of ISO 10589)

2.2. Point-to-point Adjacencies

Data center network fabrics only contain point-to-point links; because of this, there is no reason to support any broadcast link types, nor to support the Designated Intermediate System processing, including pseudonode creation. In light of this, processing related to sections 7.2.3 (broadcast networks), 7.3.8 (generation of level 1 pseudonode LSPs), 7.3.10 (generation of level 2 pseudonode LSPs), and

section 8.4.5 (LAN designated intermediate systems) in [ISO10589] SHOULD BE removed.

2.3. Three Way Handshake Support

It is important that two way connectivity be established before synchronizing the link state database, or routing through a link in a data center fabric. To reject optical failures that cause a one way connection between two routers, fabricDC must support the three way handshake mechanism described in [RFC5303].

2.4. Adjacency Formation Optimization

While adjacency formation is not considered particularly burdensome in IS-IS, it may still be useful to reduce the amount of state transferred across the network when connecting a new IS to the fabric. In its simplest form, the process is:

- o An IS connected to the fabric will send hellos on all links.
- o The IS will only complete the three-way handshake with one newly discovered neighbor; this would normally be the first neighbor which sends the newly connected intermediate system's ID back in the three-way handshake process.
- o The IS will complete its database exchange with this one newly adjacent neighbor.
- o Once this process is completed, the IS will continue processing the remaining neighbors as normal.
- o If synchronization is not achieved within twice the dead timer on the local interface, the newly connected IS will repeat this process with the second neighbor with which it forms a three-way adjacency.

This process allows each IS newly added to the fabric to exchange a full table once; a very minimal amount of information will be transferred with the remaining neighbors to reach full synchronization.

Any such optimization is bound to present a tradeoff between several factors; the mechanism described here increases the amount of time required to form adjacencies slightly in order to reduce the total state carried across the network. An alternative mechanism could provide a better balance of the amount of information carried across the network for initial synchronization and the time required to synchronize a new IS. For instance, an IS could choose to

synchronize its database with two or three adjacent intermediate systems, which could speed the synchronization process up at the cost of carrying additional data on the network. A locally determined balance between the speed of synchronization and the amount of data carried on the network can be achieved by adjusting the number of adjacent intermediate systems the newly attached IS synchronizes with.

3. Advertisement of Reachability Information

IS-IS describes the topology in two different sets of TLVs; the first describes the set of neighbors connected to an IS, the second describes the set of reachable destination connected to an IS. There are two different forms of both of these descriptions, one of which carries what are widely called narrow metrics, the other of which carries what are widely called wide metrics. In a tightly controlled data center fabric implementation, such as the ones Openfabric is designed to support, no IS that supports narrow metrics will ever be deployed or supported; hence there is no reason to support any metric type other than wide metrics.

- o The Level 2 Link State PDU (type 20 in section 9.9 of [ISO10589]) and the scoped flooding PDU (type 10 in section 3.1 of [RFC7356]) SHOULD BE the only PDU types used to carry link state information in a Openfabric implementation
- o Processing related to the Level 1 Link State PDU (type 18) MAY BE removed from Openfabric implementations (see section 9.8 of [ISO10589])
- o Neighbor reachability MUST BE carried in TLV type 22 (see section 3 of [RFC5305])
- o IPv4 reachability SHOULD BE carried in TLV type 135 (see section 4 of [RFC5305]), or TLV type 235 for multitopology implementations (see [RFC5120])
- o IPv6 reachability SHOULD BE carried in TLV type 236 (see [RFC5308]), or TLV type 237 for multitopology implemenations (see [RFC5120])
- o Processing related to the neighbor reachability TLV (type 2, see sections 9.8 and 9.9 of [ISO10589]) SHOULD BE removed
- o Processing related to the narrow metric IP reachability TLV (types 128 and 130) SHOULD BE removed

Further, if segment routing support is desired, Openfabric MAY support the Prefix Segment Identifier sub-TLV and other TLVs as required in [I-D.ietf-isis-segment-routing-extensions].

4. Determining and Advertising Location on the Fabric

The tier to which a IS is connected is useful to enable autoconfiguration of intermediate systems connected to the fabric and to reduce flooding. Once the tier of an intermediate system within the fabric has been determined, it MUST be advertised using the 4 bit Tier field described in section 3.3 of [I-D.shen-isis-spine-leaf-ext]. This section describes a method of calculating the tier number, assuming the tier numbers rise in value from the edge of the fabric.

This method begins with two of the T0 intermediate systems advertising their location in the fabric. This information can either be obtained through:

- o Two T0 intermediate systems are manually configured to advertise 0x00 in their IS reachability tier sub-TLV, indicating they are at the edge of the fabric (a ToR IS).
- o The T0 intermediate systems detect they are T0 through the presence connected hosts (i.e. through a request for address assignment or some other means). If such detection is used, and the IS determines it is located at T0, it should advertise 0x00 in its IS reachability tier sub-TLV.

If the first method is used, the two T0 routers MUST be "maximally separated" on the fabric. They must be a maximal number of hops apart, or rather they MUST NOT be connected to the same T1 device as their "upstream" towards the superspines in a 5 ary fabric.

The second method above SHOULD be used with care, as it may not be secure, and it may not work in all data center environments. For instance, if a host is mistakenly (or intentionally, as a form of attack) attached to a spine IS, or a request for address assignment is transmitted to a spine IS during the bootup phase of the device or fabric, it is possible to cause a spine IS to advertise itself as a T0. Unless the autodetection of the T0 devices is secured, the manual mechanism SHOULD BE used (configuring at least one T0 device manually).

Given the correct configuration of two T0 devices, maximally spaced on the fabric, the remaining intermediate systems calculate their tier number as follows:

- o The local IS calculates an SPT (using SPF) setting the cost of every link to 1; this effectively calculates a topology only view of the network, without considering any configured link costs
- o Ensure that at least two T0 are in the calculated SPT; otherwise abort
- o Find the furthest T0; call this node A and set LD to the cost; the "furthest T0" is the T0 with the largest metric, or the furthest distance from the local calculating node
- o Calculate an SPT (using SPF) from the perspective of A (above) setting the cost of every link to 1
- o Find the furthest IS in A's SPT; call this node B and set RD to the cost from A to B
- o Calculate the tier number of the local IS by subtracting LD from RD

In the example network, assume 5A and 1C are manually configured as a T0, and are advertising their tier numbers. From here:

- o From 1A the path to 5A is 4 hops; this is LD
- o Run SPF from the perspective of 5A with all link metrics set to 1
- o From 5A the path length to 1C is 4; this is RD
- o $RD - LD$ is 0 at 1A, so 1A is T0, or a ToR

This process will work for any spine and leaf fabric without "cross links."

5. Flooding Optimization

Flooding is perhaps the most challenging scaling issue for a link state protocol running on a dense, large scale fabric. To reduce the flooding of link state information in the form of Link State Protocol Data Units (LSPs), Openfabric takes advantage of information already available in the link state protocol, the list of the local intermediate system's neighbor's neighbors, and the fabric locality computed above. The following tables are required to compute a set of reflooders:

- o Neighbor List (NL) list: The set of neighbors

- o Neighbor's Neighbors (NN) list: The set of neighbor's neighbors; this can be calculated by running SPF truncated to two hops
- o Do Not Reflood (DNR) list: The set of neighbors who should have LSPs (or fragments) who should not reflood LSPs
- o Reflood (RF) list: The set of neighbors who should flood LSPs (or fragments) to their adjacent neighbors to ensure synchronization

NL is set to contain all neighbors, and sorted deterministically (for instance, from the highest IS identifier to the lowest). All intermediate systems within a single fabric SHOULD use the same mechanism for sorting the NL list. NN is set to contain all neighbor's neighbors, or all intermediate systems that are two hops away, as determined by performing a truncated SPF. The DNR and RF tables are initially empty. To begin, the following steps are taken to reduce the size of NN and NL:

- o Move any IS in NL with its tier (or fabric location) set to T0 to DNR
- o Remove all intermediate systems from NL and NN that in the shortest path to the IS that originated the LSP

Then, for every IS in NL:

- o If the current entry in NL is connected to any entries in NN:
 - * Move the IS to RF
 - * Remove the intermediate systems connected to the IS from NN
- o Else move the IS to DNR

The calculation terminates when the NL is empty.

When flooding, LSPs transmitted to adjacent neighbors on the RF list will be transmitted normally. Adjacent intermediate systems on this list will reflood received LSPs into the next stage of the topology, ensuring database synchronization. LSPs transmitted to adjacent neighbors on the DNR list, however, MUST be transmitted using a circuit scope PDU as described in [RFC7356].

5.1. Flooding Failures

It is possible in some failure modes for flooding to be incomplete because of the flooding optimizations outlined. Specifically, if a reflooder fails, or is somehow disconnected from all the links across

which it should be reflooding, it is possible an LSP is only partially flooded through the fabric. To prevent such situations, any IS receiving an LSP transmitted using DNR SHOULD:

- o Set a short timer; the default should be less than one second
- o When the timer expires, send a Complete Sequence Number Packet (CSNP) to all neighbors
- o Process any Partial Sequence Number Packets (PSNPs) as required to resynchronize
- o If a resynchronization is required, notify the network operator through a network management system

6. Other Optimizations

6.1. Transit Link Reachability

In order to reduce the amount of control plane state carried on large scale spine and leaf fabrics, openfabric implementations SHOULD NOT advertise reachability for transit links. These links MAY remain unnumbered, as IS-IS does not require layer 3 IP addresses to operate. Each IS SHOULD be configured with a single loopback address, which is assigned an IPv6 address, to provide reachability to intermediate systems which make up the fabric.

[RFC3277] SHOULD be supported on devices supporting openfabric with unnumbered interface in order to support traceability and network management.

6.2. Transiting T0 Intermediate Systems

In data center fabrics, ToR intermediate systems SHOULD NOT be used to transit between two T1 (or above) spine intermediate systems. The simplest way to prevent this is to set the overload bit [RFC3277] for all the LSPs originated from T0 intermediate systems. However, this solution would have the unfortunate side effect of causing all reachability beyond any T0 IS to have the same metric, and many implementations treat a set overload bit as a metric of 0xFFFF in calculating the Shortest Path Tree (SPT). This document proposes an alternate solution which preserves the leaf node metric, while still avoiding transiting T0 intermediate systems.

Specifically, all T0 intermediate systems SHOULD advertise their metric to reach any T1 adjacent neighbor with a cost of 0XFFE. T1 intermediate systems, on the other hand, will advertise T0 intermediate systems with the actual interface cost used to reach the

T0 IS. Hence, links connecting T0 and T1 intermediate systems will be advertised with an asymmetric cost that discourages transiting T0 intermediate systems, while leaving reachability to the destinations attached to T0 devices the same.

7. Openfabric and Route Aggregation

While schemes may be designed so reachability information can be aggregated in Openfabric deployments, this is not a recommended configuration.

8. Security Considerations

This document outlines modifications to the IS-IS protocol for operation on large scale data center fabrics. While it does add new TLVs, and some local processing changes, it does not add any new security vulnerabilities to the operation of IS-IS. However, openfabric implementations SHOULD implement IS-IS cryptographic authentication, as described in [RFC5304], and should enable other security measures in accordance with best common practices for the IS-IS protocol.

If T0 intermediate systems are auto-detected using information outside Openfabric, it is possible to attack the calculations used for flooding reduction and auto-configuration of intermediate systems. For instance, if a request for an address pool is used as an indicator of an attached host, and hence receiving such a request causes an intermediate system to advertise itself as T0, it is possible for an attacker (or a simple mistake) to cause auto-configuration to fail. Any such auto-detection mechanisms SHOULD BE secured using appropriate techniques, as described by any protocols or mechanisms used.

9. References

9.1. Normative References

[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-07 (work in progress), October 2018.

- [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>>.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <<https://www.rfc-editor.org/info/rfc2629>>.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <<https://www.rfc-editor.org/info/rfc5120>>.
- [RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, DOI 10.17487/RFC5301, October 2008, <<https://www.rfc-editor.org/info/rfc5301>>.
- [RFC5303] Katz, D., Saluja, R., and D. Eastlake 3rd, "Three-Way Handshake for IS-IS Point-to-Point Adjacencies", RFC 5303, DOI 10.17487/RFC5303, October 2008, <<https://www.rfc-editor.org/info/rfc5303>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, DOI 10.17487/RFC5308, October 2008, <<https://www.rfc-editor.org/info/rfc5308>>.
- [RFC5309] Shen, N., Ed. and A. Zinin, Ed., "Point-to-Point Operation over LAN in Link State Routing Protocols", RFC 5309, DOI 10.17487/RFC5309, October 2008, <<https://www.rfc-editor.org/info/rfc5309>>.

- [RFC5311] McPherson, D., Ed., Ginsberg, L., Previdi, S., and M. Shand, "Simplified Extension of Link State PDU (LSP) Space for IS-IS", RFC 5311, DOI 10.17487/RFC5311, February 2009, <<https://www.rfc-editor.org/info/rfc5311>>.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <<https://www.rfc-editor.org/info/rfc5316>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [RFC7981] Ginsberg, L., Previdi, S., and M. Chen, "IS-IS Extensions for Advertising Router Information", RFC 7981, DOI 10.17487/RFC7981, October 2016, <<https://www.rfc-editor.org/info/rfc7981>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

9.2. Informative References

- [I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-19 (work in progress), July 2018.
- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.
- [RFC3277] McPherson, D., "Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance", RFC 3277, DOI 10.17487/RFC3277, April 2002, <<https://www.rfc-editor.org/info/rfc3277>>.
- [RFC3719] Parker, J., Ed., "Recommendations for Interoperable Networks using Intermediate System to Intermediate System (IS-IS)", RFC 3719, DOI 10.17487/RFC3719, February 2004, <<https://www.rfc-editor.org/info/rfc3719>>.

- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <<https://www.rfc-editor.org/info/rfc5304>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC5449] Baccelli, E., Jacquet, P., Nguyen, D., and T. Clausen, "OSPF Multipoint Relay (MPR) Extension for Ad Hoc Networks", RFC 5449, DOI 10.17487/RFC5449, February 2009, <<https://www.rfc-editor.org/info/rfc5449>>.
- [RFC5614] Ogier, R. and P. Spagnolo, "Mobile Ad Hoc Network (MANET) Extension of OSPF Using Connected Dominating Set (CDS) Flooding", RFC 5614, DOI 10.17487/RFC5614, August 2009, <<https://www.rfc-editor.org/info/rfc5614>>.
- [RFC5820] Roy, A., Ed. and M. Chandra, Ed., "Extensions to OSPF to Support Mobile Ad Hoc Networking", RFC 5820, DOI 10.17487/RFC5820, March 2010, <<https://www.rfc-editor.org/info/rfc5820>>.
- [RFC5837] Atlas, A., Ed., Bonica, R., Ed., Pignataro, C., Ed., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, DOI 10.17487/RFC5837, April 2010, <<https://www.rfc-editor.org/info/rfc5837>>.
- [RFC6232] Wei, F., Qin, Y., Li, Z., Li, T., and J. Dong, "Purge Originator Identification TLV for IS-IS", RFC 6232, DOI 10.17487/RFC6232, May 2011, <<https://www.rfc-editor.org/info/rfc6232>>.
- [RFC7182] Herberg, U., Clausen, T., and C. Dearlove, "Integrity Check Value and Timestamp TLV Definitions for Mobile Ad Hoc Networks (MANETs)", RFC 7182, DOI 10.17487/RFC7182, April 2014, <<https://www.rfc-editor.org/info/rfc7182>>.
- [RFC7921] Atlas, A., Halpern, J., Hares, S., Ward, D., and T. Nadeau, "An Architecture for the Interface to the Routing System", RFC 7921, DOI 10.17487/RFC7921, June 2016, <<https://www.rfc-editor.org/info/rfc7921>>.

Appendix A. Flooding Optimization Operation

Recent testing has shown that flooding is largely a "non-issue" in terms of scaling when using high speed links connecting intermediate systems with reasonable processing power and memory. However, testing has also shown that flooding will impact convergence speed even in such environments, and flooding optimization has a major impact on the performance of a link state protocol in resource constrained environments. Some thoughts on flooding optimization in general, and the flooding optimization contained in this document, follow.

There are two general classes of flooding optimization available for link state protocols. The first class of optimization relies on a centralized service or server to gather the link state information and redistribute it back into the intermediate systems making up the fabric. Such solutions are attractive in many, but not all, environments; hence these systems compliment, rather than compete with, the system described here. Systems relying on a service or server necessarily also rely on connectivity to that service or server, either through an out-of-band network or connectivity through the fabric itself. Because of this, these mechanisms do not apply to all deployments; some deployments require underlying reachability regardless of connectivity to an outside service or server.

The second possibility is to create a fully distributed system that floods the minimal amount of information possible to every intermediate system. The system described in this draft is an example of such a system. Again, there are many ways to accomplish this goal, but simplicity is a primary goal of the system described in this draft.

The system described here divides the work into two different parts; forward and reverse optimization. The forward optimization begins by finding the set of intermediate systems two hops away from the flooding device, and choosing a subset of connected neighbors that will successfully reach this entire set of intermediate systems, as shown in the diagram below.

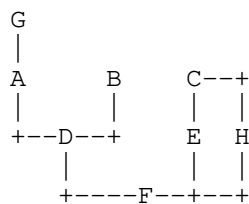


Figure 2

If F is flooding some piece of information, then it will find the entire set of intermediate systems within two hops by discovering its neighbors and their neighbors from the local LSDB. This will include A, B, C, D, and E--but not G. From this set, F can determine that D can reach A and B, while a single flood to either E or H will reach C. Hence F can flood to D and either E or H to reach C. F can choose to flood to D and E normally. Because H still needs to receive this new LSP (or fragment!), but does not need to reflood to C, F can send the LSP using link local signaling. In this case, H will receive and process the new LSP, but not reflood it.

Rather than carrying the information necessary through hello extensions, as is done in [RFC5820], the neighbors are allowed to complete initial synchronization, and then a truncated shortest path tree is built to determine the "two hop neighborhood." This has the advantage of using mechanisms already used in IS-IS, rather than adding new processes. The risk with this process is any LSPs flooded through the network before this initial calculation takes place will be suboptimal. This "two hop neighborhood" process has been used in OSPF deployments for a number of years, and has proven stable in practice.

Rather than setting a timer for reflooding, the implementation described here uses IS-IS' ability to describe the entire database using a CSNP to ensure flooding is successful. This adds some small amount of overhead, so there is some balance between optimal flooding and ensuring flooding is complete.

The reverse optimization is simpler. It relies on the observation that any intermediate system between the local IS and the origin of the LSP, other than in the case of floods removing an LSP from the shared LSDB, should have already received a copy of the LSP. For instance, if F originates an LSP in the figure above, and E refloods the LSP to C, C does not need to reflood back to F if F is on its shortest path tree towards F. It is obvious this is not a "perfect" optimization. A perfect optimization would block flooding back along a directed acyclic graph towards the originator. Using the SPT, however, is a quick way to reduce flooding without performing more calculations.

The combination of these two optimizations have been seen, in testing, to reduce the number of copies any IS receives from the tens to precisely one.

Appendix B. Fabric Location Calculation

Determining the location of a device in a symmetric topology is quite challenging. The authors of this draft worked through a number of possible solutions to this problem, each of which was found to either not work in some topology, or was found to be liable to unacceptable errors. For instance:

- o Method 1:

- * Caculate the maximum distance through the fabric, and the distance from one of those points to the local intermediate system
- * This works in a five stage Clos spine and leaf, but not in a three stage, nor in some other five stage spine and leaf fabrics, such as the common butterfly or Benes fabric

- o Method 2:

- * Manually mark one edge leaf node in the fabric as T0
- * Calculate maximum distance through the fabric from this point
- * Calculate local position based on this maximum distance the distance to the single marked device
- * This works in three and five stage Clod fabrics, but does not work from every location in other spine and leaf fabrics, such as the common butterfly or Benes fabric

In the end, marking two devices located as far from one another topologically as possible provides the anchor points necessary to calculate the total distance through the fabric, and then from those points to the location of the calculating device.

The information obtained in this way can also be combined with other forms of location calculation, such as whether a device requesting an address through some mechanism is attached to the local device, or other indications of fabric locality. It generally true that having more than one method to determine fabric location will be better than any single method to account for errors, failures, and other problems that can arise with any mechanism.

Authors' Addresses

Russ White (editor)
LinkedIn

Email: russ@riw.us

Shawn Zandi (editor)
LinkedIn

Email: szandi@linkedin.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2018

Y. Zhu
H. Chen
China Telecom
Z. Du
M. Chen
Huawei
July 3, 2017

ISIS Extensions for Flexible Ethernet
draft-zcdc-isis-flexe-extention-01

Abstract

This document specifies the extensions to the IS-IS routing protocol to carry and flood Flex Ethernet (FlexE) link state information. The FlexE link state information is necessary for a node or a controller to compute a path that is required to over FlexE links.

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 January 4, 2018.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. FlexE Link Advertisement	3
3. FlexE Sub-link Advertisement	6
4. IANA Considerations	6
4.1. FlexE Switching Type	6
4.2. FlexE LSP Encoding Type	6
4.3. FlexE Interface Sub-TLV	7
5. Security Consideration	7
6. Acknowledgements	7
7. References	7
7.1. Normative References	7
7.2. Informative References	7
Authors' Addresses	8

1. Introduction

Flex Ethernet (FlexE) [I-D.izh-ccamp-flexe-fwk] provides a generic mechanism for supporting a variety of Ethernet MAC rates that may or may not correspond to any existing Ethernet PHY rate. This includes MAC rates that are both greater than (through bonding) and less than (through sub-rate and channelization) the Ethernet PHY rates used to carry Ethernet traffic.

FlexE supports interface bonding, a bonded interface is consisted of from 1 to n 100GBASE-R PHYs (other types of PHY will be supported in the future), the bonded interface is called FlexE interface in this document. FlexE also supports interface channelization, a FlexE interface can be channelized into multiple sub-interfaces, the sub-interface is called FlexE sub-interface in the rest of this document.

The FlexE mechanism operates using a calendar which assigns 66B block positions on sub-calendars on each PHY of a FlexE interface to each of the FlexE flows. The calendar has a granularity of 5G, and has a length of 20 slots for a 100G interface. Currently, only 100GBASE-R PHY and 5G granularity are supported in FlexE implementation

agreement version 1.0 [FlexE], other types (e.g., 200G, 400G) of PHY and granularities (e.g., 25G) will be supported in the future.

A FlexE interface has a number of time slots resource. These time slots can be transparent to the up layer application, the up layer application (e.g., RSVP-TE) can just treat the FlexE interface as a normal Ethernet interface, or the time slots can be allocated to a FlexE LSP through RSVP-TE signaling, or the time slots can be allocated to form a FlexE sub-interface through configuration or some dynamic protocols. How to signal the FlexE LSP or configure the FlexE sub-interface is out of the scope of this document.

The logical link that connects two FlexE interfaces residing in two adjacent nodes is called FlexE link, and the logical link that connects two FlexE sub-interfaces residing in two adjacent nodes is called FlexE sub-link.

More details about FlexE can be found in FlexE framework document [I-D.izh-ccamp-flexe-fwk].

This document defines extensions to ISIS protocol to advertise the FlexE TE link and sub-link state information.

2. FlexE Link Advertisement

This document re-uses the Interface Switching Capability Descriptor (ISCD) sub-TLV for the advertisement of FlexE link state information. The ISCD is a sub-TLV of the extended IS reachability TLV [RFC5307], it is defined to describe the switching capability of an interface. The following figure (Figure 1) illustrates encoding of the Value field of the ISCD sub-TLV.

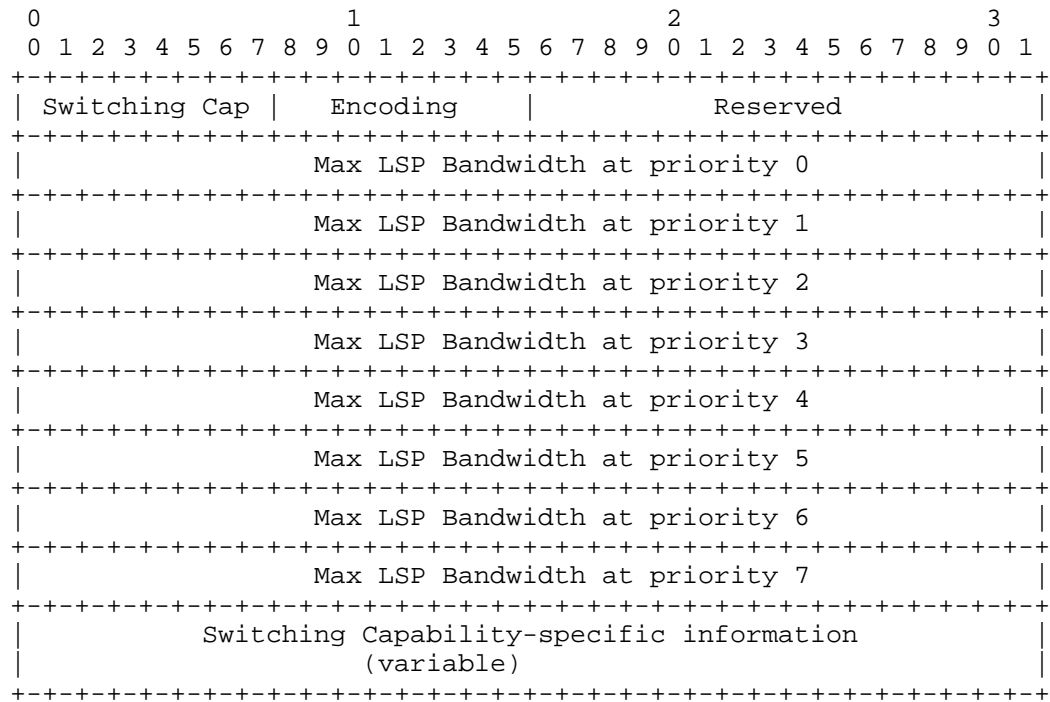


Figure 1: ISCD sub-TLV

To support FlexE link advertisement, new "Switching Cap" and "Encoding" are defined as follows:

The Switching Capability (Switching Cap) for FlexE interface is as below:

Value	Name
TBD1	FlexE-Switching

The Encoding Type for FlexE:

Value	Name
TBD2	FlexE

The "Switching Capability-specific information" field for FlexE interface is defined as below. It is referred to as FlexE Interface sub-TLV in this document.

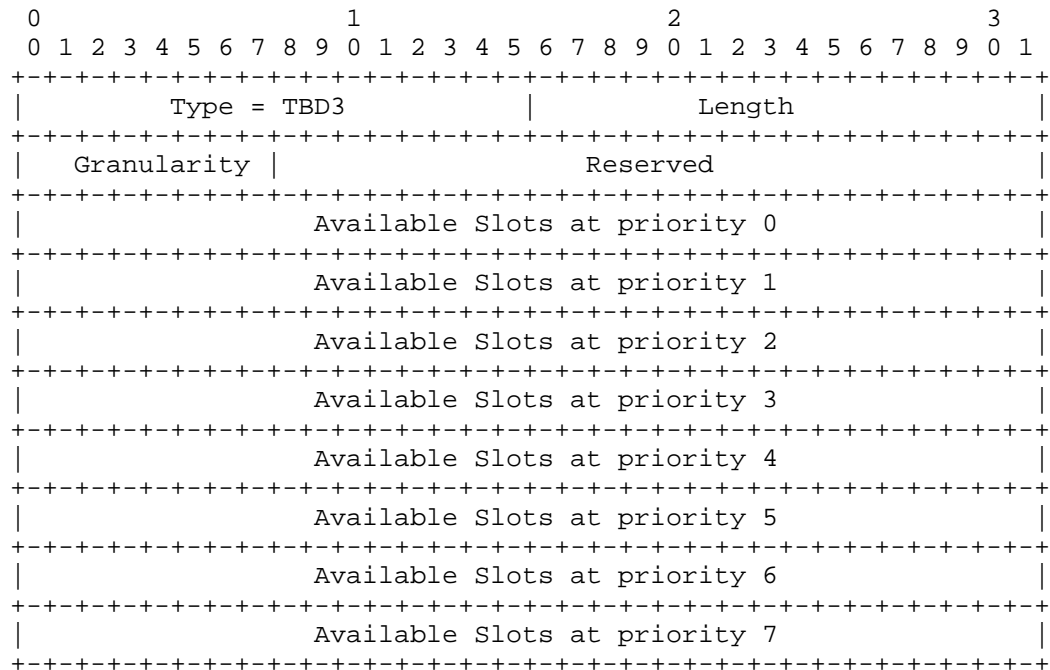


Figure 2: FlexE Interface sub-TLV

The Type field is 2 octets in length and the value is TBD3.

The Length field is 2 octets in length that indicates the total length of the TLV in octet.

The Granularity is 1 octet in length and its value identifies the granularity of the FlexE time slots of a FlexE interface. Current OIF agreement only allows the "5G" granularity, other granularities may be defined in the future.

Value	Granularity
-----	-----
0	Reserved
1	5G
2-254	Unassigned
255	Reserved

For each PHY of a FlexE interface, there are two calendars, one is called Active calendar and the other is called Backup calendar. The two calendars are used to facilitate reconfiguration, for example,

FlexE flow resizing can be achieved through calendar updates. More detail about FlexE calendar can be found [FlexE].

Each Available Slots at priority n is 4-octet in length that indicates the maximum number of slots available at priority 'n' on active calendar of the FlexE interface.

For a FlexE interface, as said above, 5G granularity is only supported for now, but multiple granularities may be supported in the future. To support this, FlexE Interface sub-TLV can occur multiple times in a ISCD sub-TLV, but for each granularity, only one FlexE Interface sub-TLV can be included and it carries the available time slots of the granularity of the FlexE interface. When multiple FlexE Interface sub-TLVs for the same granularity occur, only the first FlexE Interface sub-TLV is considered to be valid, the rests MUST be ignored.

3. FlexE Sub-link Advertisement

Through FlexE channelization, a FlexE Link can be sliced into a number of FlexE sub-links, each FlexE sub-link has dedicated bandwidth and is isolated from other FlexE sub-links. A set of FlexE sub-links can be allocated to a specific application/user to form a sliced network. From link characteristic point of view, a FlexE sub-link is same as a real point-2-point link, it can be advertised and used as a normal point-2-point link.

4. IANA Considerations

4.1. FlexE Switching Type

IANA is requested to allocate a new switching type from the "Switching Types" registry of "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Parameters" registry.

Value	Name	Reference
-----	-----	-----
TBD1	FlexE-Switching	This document

4.2. FlexE LSP Encoding Type

IANA is requested to allocate a new LSP encoding type from the "LSP Encoding Types" registry of "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Parameters" registry.

Value	Name	Reference
-----	-----	-----
TBD2	FlexE	This document

4.3. FlexE Interface Sub-TLV

IANA is requested to create and maintain a new sub-registry, the "Types for sub-TLVs of FlexE Switching Capability Specific Information" registry under the "IS-IS TLV Codepoints" registry.

Value	sub-TLV Name	Reference
-----	-----	-----
TBD3	FlexE Interface	This document

5. Security Consideration

This document describes a mechanism for advertising FlexE link state information through IS-IS LSPs and does not introduce any new security issues.

6. Acknowledgements

7. References

7.1. 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>>.
- [RFC5029] Vasseur, JP. and S. Previdi, "Definition of an IS-IS Link Attribute Sub-TLV", RFC 5029, DOI 10.17487/RFC5029, September 2007, <<http://www.rfc-editor.org/info/rfc5029>>.
- [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>>.

7.2. Informative References

- [FlexE] OIF, "Flex Ethernet Implementation Agreement Version 1.0 (OIF-FLEXE-01.0)", March 2016.
- [I-D.izh-ccamp-flex-e-fwk] Hussain, I., Valiveti, R., Wang, Q., Andersson, L., Chen, M., and z. zhenghaomian@huawei.com, "GMPLS Routing and Signaling Framework for Flexible Ethernet (FlexE)", draft-izh-ccamp-flex-e-fwk-03 (work in progress), June 2017.

Authors' Addresses

Yongqing Zhu
China Telecom
109, West Zhongshan Road, Tianhe District, Guangzhou, China

Email: zhuyq@gsta.com

Huanan Chen
China Telecom
109, West Zhongshan Road, Tianhe District, Guangzhou, China

Email: chenhuanan@gsta.com

Zongpeng Du
Huawei

Email: duzongpeng@huawei.com

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com