LSR Workgroup A. Lindem
Internet-Draft Cisco Systems
Intended status: Standards Track Y. Qu
Expires: August 23, 2021 Futurewei
A. Roy
Arrcus, Inc.
S. Mirtorabi
Cisco Systems
February 19, 2021

OSPF Transport Instance Extensions
draft-acee-lsr-ospf-transport-instance-02

Abstract

OSPFv2 and OSPFv3 include a reliable flooding mechanism to disseminate routing topology and Traffic Engineering (TE) information within a routing domain. Given the effectiveness of these mechanisms, it is convenient to envision using the same mechanism for dissemination of other types of information within the domain. However, burdening OSPF with this additional information will impact intra-domain routing convergence and possibly jeopardize the stability of the OSPF routing domain. This document presents mechanism to relegate this ancillary information to a separate OSPF instance and minimize the impact.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

OSPFv2 [RFC2328] and OSPFv3 [RFC5340] include a reliable flooding mechanism to disseminate routing topology and Traffic Engineering (TE) information within a routing domain. Given the effectiveness of these mechanisms, it is convenient to envision using the same mechanism for dissemination of other types of information within the domain. However, burdening OSPF with this additional information will impact intra-domain routing convergence and possibly jeopardize the stability of the OSPF routing domain. This document presents mechanism to relegate this ancillary information to a separate OSPF instance and minimize the impact.

This OSPF protocol extension provides functionality similar to "Advertising Generic Information in IS-IS" [RFC6823]. Additionally, OSPF is extended to support sparse non-routing overlay topologies Section 4.7.

2. Requirements Language

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

3. Possible Use Cases

3.1. MEC Service Discovery

Multi-Access Edge Computing (MEC) plays an important role in 5G architecture. MEC optimizes the performance for ultra-low latency and high bandwidth services by providing networking and computing at the edge of the network [ETSI-WP28-MEC]. To achieve this goal, it's important to expose the network capabilities and services of a MEC device to 5G User Equipment UE, i.e. UEs.

The followings are an incomplete list of the kind of information that OSPF transport instance can help to disseminate:

- o A network service is realized using one or more physical or virtualized hosts in MEC, and the locations of these service points might change. The auto-discovery of these service locations can be achieved using an OSPF transport instance.

- o UEs might be mobile, and MEC should support service continuity and application mobility. This may require service state transferring and synchronization. OSPF transport instance can be used to synchronize these states.

- o Network resources are limited, such as computing power, storage. The availability of such resources is dynamic, and OSPF transport instance can be used to populate such information, so applications can pick the right location of such resources, hence improve user experience and resource utilization.

3.2. Application Data Dissemination

Typically a network consists of routers from different vendors with different capabilities, and some applications may want to know whether a router supports certain functionality or where to find a router supports a functionality, so it will be ideal if such kind of information is known to all routers or a group of routers in the network. For example, an ingress router needs to find an egress router that supports In-situ Flow Information Telemetry (IFIT) [I-D.wang-lsr-igp-extensions-ifit] and obtain IFIT parameters.

OSPF transport instance can be used to populate such router capabilities/functionalities without impacting the performance or convergence of the base OSPF protocol.

3.3. Intra-Area Topology for BGP-LS Distribution

In some cases, it is desirable to limit the number of BGP-LS [RFC5572] sessions with a controller to the a one or two routers in an OSPF domain. However, many times those router(s) do not have full visibility to the complete topology of all the areas. To solve this problem without extended the BGP-LS domain, the OSPF LSAs for non-local area could be flooded over the OSPF transport instance topology using remote neighbors Section 4.7.1.

4. OSPF Transport Instance

In order to isolate the effects of flooding and processing of non-routing information, it will be relegated to a separate protocol instance. This instance should be given lower priority when contending for router resources including processing, backplane bandwidth, and line card bandwidth. How that is realized is an implementation issue and is outside the scope of this document.

Throughout the document, non-routing refers to routing information that is not used for IP or IPv6 routing calculations. The OSPF transport instance is ideally suited for dissemination of routing information for other protocols and layers.

4.1. OSPFv2 Transport Instance Packet Differentiation

OSPFv2 currently does not offer a mechanism to differentiate Transport instance packets from normal instance packets sent and received on the same interface. However, the [RFC6549] provides the necessary packet encoding to support multiple OSPF protocol instances.

4.2. OSPFv3 Transport Instance Packet Differentiation

Fortunately, OSPFv3 already supports separate instances within the packet encodings. The existing OSPFv3 packet header instance ID field will be used to differentiate packets received on the same link (refer to section 2.4 in [RFC5340]).

4.3. Instance Relationship to Normal OSPF Instances

In OSPF transport instance, we must guarantee that any information we've received is treated as valid if and only if the router sending it is reachable. We'll refer to this as the "condition of reachability" in this document.

The OSPF transport instance is not dependent on any other OSPF instance. It does, however, have much of the same as topology information must be advertised to satisfy the "condition of reachability".

Further optimizations and coupling between an OSPF transport instance and a normal OSPF instance are beyond the scope of this document. This is an area for future study.

4.4. Network Prioritization

While OSPFv2 (section 4.3 in [RFC2328]) are normally sent with IP precedence Internetwork Control, any packets sent by an OSPF transport instance will be sent with IP precedence Flash (B'011'). This is only appropriate given that this is a pretty flashy mechanism.

Similarly, OSPFv3 transport instance packets will be sent with the traffic class mapped to flash (B'011') as specified in ([RFC5340]).

By setting the IP/IPv6 precedence differently for OSPF transport instance packets, normal OSPF routing instances can be given priority during both packet transmission and reception. In fact, some router implementations map the IP precedence directly to their internal packet priority. However, internal router implementation decisions are beyond the scope of this document.

4.5. OSPF Transport Instance Omission of Routing Calculation

Since the whole point of the transport instance is to separate the routing and non-routing processing and fate sharing, a transport instance SHOULD NOT install any IP or IPv6 routes. OSPF routers SHOULD NOT advertise any transport instance LSAs containing IP or IPv6 prefixes and OSPF routers receiving LSAs advertising IP or IPv6 prefixes SHOULD ignore them. This implies that an OSPF transport instance Link State Database should not include any of the LSAs as shown in Table 1.

OSPFv2	summary-LSAs (type 3) AS-external-LSAs (type 5) NSSA-LSAs (type 7)
OSPFv3	inter-area-prefix-LSAs (type 2003) AS-external-LSAs (type 0x4005) NSSA-LSAs (type 0x2007) intra-area-prefix-LSAs (type 0x2009)
OSPFv3 Extended LSA	E-inter-area-prefix-LSAs (type 0xA023) E-as-external-LSAs (type 0xC025) E-Type-7-NSSA (type 0xA027) E-intra-area-prefix-LSA (type 0xA029)

LSAs not included in OSPF transport instance

If these LSAs are erroneously advertised, they will be flooded as per standard OSPF but MUST be ignored by OSPF routers supporting this specification.

4.6. Non-routing Instance Separation

It has been suggested that an implementation could obtain the same level of separation between IP routing information and non-routing information in a single instance with slight modifications to the OSPF protocol. The authors refute this contention for the following reasons:

o Adding internal and external mechanisms to prioritize routing information over non-routing information are much more complex than simply relegating the non-routing information to a separate instance as proposed in this specification.

- o The instance boundary offers much better separation for allocation of finite resources such as buffers, memory, processor cores, sockets, and bandwidth.

- o The instance boundary decreases the level of fate sharing for failures. Each instance may be implemented as a separate process or task.

- o With non-routing information, many times not every router in the OSPF routing domain requires knowledge of every piece of non-routing information. In these cases, groups of routers which need to share information can be segregated into sparse topologies greatly reducing the amount of non-routing information any single router needs to maintain.

4.7. Non-Routing Sparse Topologies

With non-routing information, many times not every router in the OSPF routing domain requires knowledge of every piece of non-routing information. In these cases, groups of routers which need to share information can be segregated into sparse topologies. This will greatly reduce the amount of information any single router needs to maintain with the core routers possibly not requiring any non-routing information at all.

With normal OSPF, every router in an OSPF area must have every piece of topological information and every intra-area IP or IPv6 prefix. With non-routing information, only the routers needing to share a set of information need be part of the corresponding sparse topology. For directly attached routers, one only needs to configure the desired topologies on the interfaces with routers requiring the non-routing information. When the routers making up the sparse topology are not part of a uniconnected graph, two alternatives exist. The first alternative is configure tunnels to form a fully connected graph including only those routers in the sparse topology. The second alternative is use remote neighbors as described in Section 4.7.1.

4.7.1. Remote OSPF Neighbor

With sparse topologies, OSPF routers sharing non-routing information may not be directly connected. OSPF adjacencies with remote neighbors are formed exactly as they are with regular OSPF neighbors. The main difference is that a remote OSPF neighbor's address is configured and IP routing is used to deliver OSPF protocol packets to the remote neighbor. Other salient feature of the remote neighbor include:

- o All OSPF packets have the remote neighbor's configured IP address as the IP destination address.

- o The adjacency is represented in the router Router-LSA as a router (type-1) link with the link data set to the remote neighbor's configured IP address.

- o Similar to NBMA networks, a poll-interval is configured to determine if the remote neighbor is reachable. This value is normally much higher than the hello interval with 40 seconds RECOMMENDED as the default.

4.8. Multiple Topologies

For some applications, the information need to be flooded only to a topology which is a subset of routers of the transport instance. This allows the application specific information only to be flooded to routers that support the application. A transport instance may support multiple topologies as defined in [RFC4915]. But as pointed out in Section 4.5, a transport instance or topology SHOULD NOT install any IP or IPv6 routes.

Each topology associated with the transport instance MUST be fully connected in order for the LSAs to be successfully flooded to all routers in the topology.

5. OSPF Transport Instance Information (TII) Encoding

5.1. OSPFv2 Transport Instance Information Encoding

Application specific information will be flooded in opaque LSAs as specified in [RFC5250]. An Opaque LSA option code will be reserved for Transport Instance Information (TII) as described in Section 8. The TII LSA can be advertised at any of the defined flooding scopes (link, area, or autonomous system (AS)).

```
        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
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |            LS age             |     Options   |  9, 10, or 11 |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |       TBD1    |      Opaque ID (Instance ID)                  |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                      Advertising Router                       |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                      LS sequence number                       |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |         LS checksum           |             length            |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                                                               |
       +-                            TLVs                             -+
       |                             ...                               |
```

g

OSPFv2 Transport Instance Information Opaque LSA

The format of the TLVs within the body of an TII LSA is as defined in Section 5.3.

5.2. OSPFv3 Transport Instance Information Encoding

Application specific information will be flooded in separate LSAs with a separate function code. Refer to section A.4.2.1 of [RFC5340]. for information on the LS Type encoding in OSPFv3, and section 2 of [RFC8362] for OSPFv3 extended LSA types. An OSPFv3 function code will be reserved for Transport Instance Information (TII) as described in Section 8. Same as OSPFv2, the TII LSA can be advertised at any of the defined flooding scopes (link, area, or autonomous system (AS)). The U bit will be set indicating that OSPFv3 TTI LSAs should be flooded even if it is not understood.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            LS age             |1|S12|          TBD2           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Link State ID (Instance ID)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Advertising Router                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       LS sequence number                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        LS checksum            |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+-                            TLVs                             -+
|                             ...                               |
```

OSPFv3 Transport Instance Information LSA

The format of the TLVs within the body of an TII LSA is as defined in Section 5.3.

5.3. Transport Instance Information (TII) TLV Encoding

The format of the TLVs within the body of the LSAs containing non-routing information is the same as the format used by the Traffic Engineering Extensions to OSPF [RFC3630]. The LSA payload consists of one or more nested Type/Length/Value (TLV) triplets. The format of each TLV is:

```
 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            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Value...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

TLV Format

5.3.1. Top-Level TII Application TLV

An Application top-level TLV will be used to encapsulate application data advertised within TII LSAs. This top-level TLV may be used to handle the local publication/subscription for application specific data. The details of such a publication/subscription mechanism are beyond the scope of this document. An Application ID is used in the top-level application TLV and shares the same code point with IS-IS as defined in [RFC6823].

```
 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 (1)         |        Length - Variable      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Application ID             |         Reserved              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
.                            Sub-TLVs                           .
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Application ID:
An identifier assigned to this application via the IANA registry, as defined in RFC 6823. Each unique application will have a unique ID.

Additional Application-Specific Sub-TLVs:
Additional information defined by applications can be encoded as Sub-TLVs. Definition of such information is beyond the scope of this document.

Top-Level TLV

The specific TLVs and sub-TLVs relating to a given application and the corresponding IANA considerations MUST be specified in the document corresponding to that application.

6. Manageability Considerations

7. Security Considerations

The security considerations for the Transport Instance will not be different for those for OSPFv2 [RFC2328] and OSPFv3 [RFC5340].

8. IANA Considerations

8.1. OSPFv2 Opaque LSA Type Assignment

IANA is requested to assign an option type, TBD1, for Transport Instance Information (TII) LSA from the "Opaque Link-State Advertisements (LSA) Option Types" registry.

8.2. OSPFv3 LSA Function Code Assignment

IANA is requested to assign a function code, TBD2, for Transport Instance Information (TII) LSAs from the "OSPFv3 LSA Function Codes" registry.

8.3. OSPF Transport Instance Information Top-Level TLV Registry

IANA is requested to create a registry for OSPF Transport Instance Information (TII) Top-Level TLVs. The first available TLV (1) is assigned to the Application TLV Section 5.3.1. The allocation of the unsigned 16-bit TLV type are defined in the table below.

Range	Assignment Policy
0	Reserved (Not to be assigned)
1	Application TLV
2-16383	Unassigned (IETF Review)
16383-32767	Unassigned (FCFS)
32768-32777	Experimentation (No assignements)
32778-65535	Reserved (Not to be assigned)

TII Top-Level TLV Registry Assignments

9. Acknowledgement

The authors would like to thank Les Ginsberg for review and comments.

10. References

10.1. Normative References

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

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <https://www.rfc-editor.org/info/rfc5250>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <https://www.rfc-editor.org/info/rfc6549>.

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

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

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

10.2. Informative References

[ETSI-WP28-MEC]
Sami Kekki, etc., "MEC in 5G Networks", 2018, <https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf>.

[I-D.wang-lsr-igp-extensions-ifit]
Wang, Y., Zhou, T., Qin, F., Chen, H., and R. Pang, "IGP Extensions for In-situ Flow Information Telemetry (IFIT) Capability Advertisement", draft-wang-lsr-igp-extensions-ifit-01 (work in progress), July 2020.

[RFC5572] Blanchet, M. and F. Parent, "IPv6 Tunnel Broker with the Tunnel Setup Protocol (TSP)", RFC 5572, DOI 10.17487/RFC5572, February 2010, <https://www.rfc-editor.org/info/rfc5572>.

Authors' Addresses

Acee Lindem
Cisco Systems
301 Midenhall Way
CARY, NC 27513
UNITED STATES

Email: acee@cisco.com

Yingzhen Qu
Futurewei
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: yingzhen.qu@futurewei.com

Abhay Roy
Arrcus, Inc.

Email: abhay@arrcus.com

Sina Mirtorabi
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
USA

Email: smirtora@cisco.com

LSR Working Group T. Saad
Internet-Draft V. Beeram
Intended status: Standards Track Juniper Networks
Expires: August 26, 2021 R. Chen
S. Peng
ZTE Corporation
B. Wen
Comcast
D. Ceccarelli
Ericsson
February 22, 2021

IGP Extensions for SR Slice Aggregate SIDs
draft-bestbar-lsr-spring-sa-00

Abstract

Segment Routing (SR) defines a set of topological "segments" within an IGP topology to enable steering over a specific SR path. These segments are advertised by the link-state routing protocols (IS-IS and OSPF).

This document describes extensions to the IS-IS that enable advertising Slice Aggregate SR segments that share the same IGP computed forwarding path but offer a forwarding treatment (e.g. scheduling and drop policy) that is associated with a specific Slice Aggregate.

Status of This Memo

Table of Contents

1. Introduction

The Segment Routing (SR) architecture [RFC8402] defines a set of topological "segments" within an IGP topology as means to enable steering over a specific SR end-to-end path. These segments are advertised by the IGP link-state routing protocols (IS-IS and OSPF). The SR control plane can be applied to both IPv6 and MPLS data planes.

The definition of a network slice for use within the IETF and the characteristics of IETF network slice are specified in [I-D.ietf-teas-ietf-network-slice-definition]. A framework for reusing IETF VPN and traffic-engineering technologies to realize IETF network slices is discussed in [I-D.nsdt-teas-ns-framework].

[I-D.bestbar-teas-ns-packet] introduces the notion of a Slice Aggregate as the construct that comprises of one of more IETF network slice traffic streams. A slice policy can be used to realize a slice aggregate by instantiating specific control and data plane resources on select topological elements in an IP/MPLS network.

[I-D.bestbar-spring-scalable-ns] describes an approach to extend SR to advertiser new SID types called Slice Aggregate (SA) SIDs. Such SA SIDs are used on a router to define the forwarding action for a packet (next-hop selection), as well as enforce the specific treatment (scheduling and drop policy) associated with the Slice Aggregate.

This document defines the IS-IS and OSPF encodings for the IGP-Prefix Segment, the IGP-Adjacency Segment, the IGP-LAN-Adjacency Segment that are required to support the signaling of SR Slice Aggregate SIDs operating over SR-MPLS and SRv6 dataplanes. When the Slice Aggregate segments have the same topology (and Algorithm for Prefix-SIDs), the SA SIDs share the same forwarding path (IGP next-hop(s)), but are associated with different forwarding treatment (e.g. scheduling and drop policy) that is associated with the specific Slice Aggregate.

2. Requirements Language

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

3. Slice Aggregate SIDs for SR-MPLS

Segment Routing can be directly instantiated on the MPLS data plane through the use of the Segment Routing header instantiated as a stack of MPLS labels defined in [RFC8402].

3.1. IS-IS Slice Aggregate Prefix-SID Sub-TLV

[RFC8667] defines the IS-IS Prefix Segment Identifier sub-TLV (Prefix-SID sub-TLV) that is applicable to SR-MPLS dataplane. The Prefix-SID sub-TLV carries the Segment Routing IGP-Prefix-SID, and is associated with a prefix advertised by a router.

A new IS-IS SR Slice Aggregate Prefix-SID (SA Prefix-SID) sub-TLV is defined to allow a router advertising a prefix to associate multiple SA Prefix-SIDs to the same prefix. The SA Prefix-SIDs associated with the same prefix share the same IGP path to the destination prefix within the specific mapped or customized topology/algorithm but offer the specific QoS treatment associated with the specific Slice Aggregate.

The Slice Aggregate ID is carried in the SA Prefix-SID sub-TLV to associate it to Prefix-SID with a specific Slice Aggregate. The SA Prefix-SID sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type=TBD1   |    Length     |     Flag      |   Algorithm   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             SA-ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     SID/Index/Label(Variable)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: SA Prefix-SID sub-TLV for SR-MPLS.

where:

Type: TBD1 (Suggested value to be assigned by IANA)

Length: Variable. Depending on the size of the SID.

The "Flags" and "SID/Index/Label" fields are the same as the Prefix-SID sub-TLV [RFC8667].

Algorithm: 1 octet. Associated algorithm. Algorithm values are defined in the IGP Algorithm Type registry

SA-ID: Identifies a specific Slice Aggregate within the IGP domain.

This sub-TLV MAY be present in any of the following TLVs:

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

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

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

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

This sub-TLV MAY appear multiple times in each TLV.

3.2. IS-IS Slice Aggregate Adjacency-SID Sub-TLV

[RFC8667] defines the IS-IS Adjacency Segment Identifier sub-TLV (Adj-SID sub-TLV). The Adj-SID sub-TLV is an optional sub-TLV carrying the Segment Routing IGP Adjacency-SID as defined in [RFC8402].

A new SR Slice Aggregate Adjacency-SID (SA Adj-SID) sub-TLV is defined to allow a router to allocate and advertise multiple SA Adj-SIDs towards the same IS-IS neighbor (adjacency). The SA Adj-SIDs allows a router to enforce the specific treatment associated with the Slice Aggregate.

The Slice Aggregate ID is carried in the SA Adj-SID sub-TLV to associate it to the specific Slice Aggregate. The SA Adj-SID sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |     Length    |     Flags     |     Weight    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             SA-ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   SID/Index/Label(Variable)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: SA Adj-SID sub-TLV for SR-MPLS.

where:

Type: TBD2 (Suggested value to be assigned by IANA)

Length: Variable. Depending on the size of the SID.

The "Flags" and "SID/Index/Label" fields are the same as the Adj-SID sub-TLV [RFC8667].

SA-ID: Identifies a specific Slice Aggregate within the IGP domain.

This sub-TLV MAY be present in any of the following TLVs:

TLV-22 (Extended IS reachability) [RFC5305].

TLV-222 (Multitopology IS) [RFC5120].

TLV-23 (IS Neighbor Attribute) [RFC5311].

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

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

Multiple Adj-SID sub-TLVs MAY be associated with a single IS-IS neighbor. This sub-TLV MAY appear multiple times in each TLV.

3.3. IS-IS Slice Aggregate LAN Adjacency-SIDs

In LAN subnetworks, [RFC8667] defines the SR-MPLS LAN-Adj-SID sub-TLV for a router to advertise the Adj-SID of each of its neighbors.

A new SR Slice Aggregate LAN Adjacency-SID (SA LAN-Adj-SID) sub-TLV is defined to allow a router to allocate and advertise multiple SA LAN-Adj-SIDs towards each of its neighbors on the LAN. The SA LAN-Adj-SIDs allows a router to enforce the specific treatment associated with the specific Slice Aggregate towards a neighbor.

The Slice Aggregate ID is carried in the SA LAN-Adj-SID sub-TLV to associate it to the specific Slice Aggregate. The SA LAN-Adj-SID sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type=TBD3   |     Length    |     Flags     |    Weight     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             SA-ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Neighbor System-ID (ID length octets)        |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   SID/Label/Index (variable)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: SA LAN Adj-SID sub-TLV for SR-MPLS.

where:

Type: TBD3 (Suggested value to be assigned by IANA)

Length: Variable. Depending on the size of the SID.

The "Flags" and "SID/Index/Label" fields are the same as the LAN-Adj-SID sub-TLV [RFC8667].

SA-ID: Identifies a specific Slice Aggregate within the IGP domain.

This sub-TLV MAY be present in any of the following TLVs:

TLV-22 (Extended IS reachability) [RFC5305].

TLV-222 (Multitopology IS) [RFC5120].

TLV-23 (IS Neighbor Attribute) [RFC5311].

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

Multiple LAN-Adj-SID sub-TLVs MAY be associated with a single IS-IS neighbor. This sub-TLV MAY appear multiple times in each TLV.

Editor Note: the OSPF Sub-TLV sections will be populated in further update.

4. Slice Aggregate SIDs for SRv6

Segment Routing can be directly instantiated on the IPv6 data plane through the use of the Segment Routing Header defined in [RFC8754]. SRv6 refers to this SR instantiation on the IPv6 dataplane.

The SRv6 Locator TLV was introduced in [I-D.ietf-lsr-isis-srv6-extensions] to advertise SRv6 Locators and End SIDs associated with each locator.

4.1. SRv6 SID Slice Aggregate Sub-Sub-TLV

The SRv6 End SID sub-TLV was introduced in [I-D.ietf-lsr-isis-srv6-extensions] to advertise SRv6 Segment Identifiers (SID) with Endpoint behaviors which do not require a particular neighbor.

The SRv6 End SID sub-TLV is advertised in the SRv6 Locator TLV, and inherits the topology/algorithm from the parent locator. The SRv6 End SID sub-TLV defined in [I-D.ietf-lsr-isis-srv6-extensions] carries optional sub-sub-TLVs.

A new SRv6 Slice Aggregate (SA) SID Sub-Sub-TLV is defined to allow a router to assign and advertise an SRv6 End SID that is associated with a specific Slice Aggregate. The SRv6 SID SA Sub-Sub-TLV allows routers to infer and enforce the specific treatment associated with the Slice Aggregate on the selected next-hops along the path to the End SID destination.

```
 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=TBD4   |     Length    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             SA-ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: SRv6 SID SA Sub-Sub-TLV format for SRv6.

where:

Type: TBD4

Length: 4 octets.

SA-ID: Identifies a specific Slice Aggregate within the IGP domain.

ISIS SRv6 SID SA Sub-Sub-TLV MUST NOT appear more than once in its parent Sub-TLV. If it appears more than once in its parent Sub- TLV, the parent Sub-TLV MUST be ignored by the receiver.

The new SRv6 SID SA Sub-Sub-TLV is an optional Sub-Sub-TLV of:

SRv6 End SID Sub-TLV (Section 7.2 of [I-D.ietf-lsr-isis-srv6-extensions])

SRv6 End.X SID Sub-TLV (Section 8.1 of [I-D.ietf-lsr-isis-srv6-extensions])

SRv6 LAN End.X SID Sub-TLV (Section 8.2 of [I-D.ietf-lsr-isis-srv6-extensions])

5. IANA Considerations

This document requests allocation for the following Sub-TLVs.

5.1. SR Slice Aggregate Prefix-SID sub-TLV

This TLV shares sub-TLV space with existing "Sub-TLVs for TLVs 135,235,226 and 237 registry".

Type: TBD1 (to be assigned by IANA).

5.2. SR Slice Aggregate Adjacency-SID sub-TLV

This TLV shares sub-TLV space with existing "Sub-TLVs for TLVs 22, 222, 23, 223 and 141 registry".

Type: TBD2 (to be assigned by IANA).

5.3. SR Slice Aggregate LAN-Adj-SID sub-TLV

This TLV shares sub-TLV space with existing "Sub-TLVs for TLVs 22, 222, 23, and 223 registry".

Type: TBD3 (to be assigned by IANA).

5.4. SRv6 SID Slice Aggregate Sub-Sub-TLV

Type: TBD4 (to be assigned by IANA).

6. Security Considerations

TBD.

7. Acknowledgement

The authors would like to thank Swamy SRK, and Prabhu Raj Villadathu Karunakaran for their review of this document, and for providing valuable feedback on it.

8. Contributors

The following individuals contributed to this document:

Colby Barth
Juniper Networks
Email: cbarth@juniper.net

Srihari R. Sangli
Juniper Networks
Email: ssangli@juniper.net

Chandra Ramachandran
Juniper Networks
Email: csekar@juniper.net

9. References

9.1. Normative References

[I-D.bestbar-spring-scalable-ns]
Saad, T. and V. Beeram, "Scalable Network Slicing over SR Networks", draft-bestbar-spring-scalable-ns-00 (work in progress), December 2020.

[I-D.bestbar-teas-ns-packet]
Saad, T., Beeram, V., Wen, B., Ceccarelli, D., Halpern, J., Peng, S., Chen, R., and X. Liu, "Realizing Network Slices in IP/MPLS Networks", draft-bestbar-teas-ns-packet-01 (work in progress), December 2020.

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extension to Support Segment Routing over IPv6 Dataplane", draft-ietf-lsr-isis-srv6-extensions-11 (work in progress), October 2020.

[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>.

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

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

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

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

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

9.2. Informative References

[I-D.ietf-teas-ietf-network-slice-definition] Rokui, R., Homma, S., Makhijani, K., Contreras, L., and J. Tantsura, "Definition of IETF Network Slices", draft-ietf-teas-ietf-network-slice-definition-00 (work in progress), January 2021.

[I-D.nsdt-teas-ns-framework] Gray, E. and J. Drake, "Framework for Transport Network Slices", draft-nsdt-teas-ns-framework-04 (work in progress), July 2020.

Authors' Addresses

Tarek Saad
Juniper Networks

Email: tsaad@juniper.net

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

Ran Chen
ZTE Corporation

Email: chen.ran@zte.com.cn

Shaofu Peng
ZTE Corporation

Email: peng.shaofu@zte.com.cn

Bin Wen
Comcast

Email: Bin_Wen@cable.comcast.com

Daniele Ceccarelli
Ericsson

Email: daniele.ceccarelli@ericsson.com

Network Working Group L. Dunbar
Internet Draft H. Chen
Intended status: Standard Futurewei
Expires: September 10, 2021 Aijun Wang
China Telecom
March 10, 2021

OSPF extension for 5G Edge Computing Service
draft-dunbar-lsr-5g-edge-compute-ospf-ext-04

Abstract

This draft describes an OSPF extension for routers to advertise the running status and environment of the directly attached 5G Edge Computing servers. The AppMetaData can be used by the routers in the 5G Local Data Network to make intelligent decisions to optimize the forwarding of flows from UEs. The goal is to improve latency and performance for 5G Edge Computing services.

Table of Contents

1. Introduction

This document describes an OSPF extension to distribute the 5G Edge Computing App running status and environment so that other routers in the 5G Local Data Network (LDN) can make intelligent decisions to optimize the forwarding of flows from UEs. The goal is to improve latency and performance for 5G Edge Computing services.

1.1. 5G Edge Computing Background

As described in [3GPP-EdgeComputing], it is desirable for a mission critical Application to have multiple Application Servers hosted in multiple Edge Computing data centers to minimize the latency and to optimize the user experience. Those Edge Computing data centers are usually very close to or co-located with 5G base stations.

When a UE (User Equipment) initiates application packets using the destination address from a DNS reply or its cache, the packets from the UE are carried in a PDU session through 5G Core [5GC] to the 5G UPF-PSA (User Plan Function - PDU Session Anchor). The UPF-PSA decapsulates the 5G GTP outer header and forwards the packets from the UEs to the Ingress router of the Edge Computing (EC) Local Data Network (LDN) which is responsible for forwarding the packets to the intended destinations.

When the UE moves out of coverage of its current gNB (next-generation Node B) (gNB1), the handover procedure is initiated which includes the 5G SMF (Session Management

Function) selecting a new UPF-PSA [3GPP TS 23.501 and TS 23.502]. When the handover process is complete, the UE has a new IP address and the IP point of attachment is to the new UPF-PSA. 5GC may maintain a path from the old UPF to new the UPF for a short time for SSC [Session and Service Continuity] mode 3 to make the handover process more seamless.

```
   +--+
   |UE|---\+---------+                 +------------------+
   +--+    |  5G     |    +---------+  |   S1: aa08::4450 |
   +--+    | Site +--++---+         +----+                |
   |UE|----|  A   |PSA| Ra|         | R1 | S2: aa08::4460 |
   +--+    |      +---+---+         +----+                |
  +---+    |         |  |           |  |   S3: aa08::4470 |
  |UE1|---/+---------+  |           |  +------------------+
  +---+                 |IP Network |       L-DN1
                        |(3GPP N6)  |
     |                  |           |  +------------------+
     | UE1              |           |  |  S1: aa08::4450   |
     | moves to         |          +----+                 |
     | Site B           |          | R3 | S2: aa08::4460   |
     v                  |          +----+                 |
                        |           |  |  S3: aa08::4470   |
                        |           |  +------------------+
                        |           |      L-DN3
   +--+                 |           |
   |UE|---\+---------+  |           |  +------------------+
   +--+    |  5G     |  |           |  |  S1: aa08::4450   |
   +--+    | Site +--++-+--+        +----+                |
   |UE|----|  B   |PSA| Rb |        | R2 | S2: aa08::4460 |
   +--+    |      +--++----+        +----+                |
   +--+    |         |  +-----------+  |  S3: aa08::4470   |
   |UE|---/+---------+                 +------------------+
   +--+                                     L-DN2
```

Figure 1: App Servers in different edge DCs

1.2. Problem#1: ANYCAST in 5G EC Environment

Increasingly, ANYCAST is used extensively by various application providers and CDNs because ANYCAST makes it possible to dynamically load balance across server locations based on network conditions. With multiple servers having the same ANYCAST address, it eliminates the single point of failure and bottleneck at the application layer load balancers. Another benefit of using ANYCAST address is removing the dependency on how UEs get the IP addresses for their Applications. Some UEs (or clients) might use stale cached IP addresses for an extended period.

But, having multiple locations of the same ANYCAST address in 5G Edge Computing environment can be problematic because all those edge computing Data Centers can be close in proximity. There might be very little difference in the routing cost to reach the Application Servers in different Edge DCs, which can cause packets from one flow to be forwarded to different locations, resulting in service glitches.

1.3. Problem #2: Unbalanced Anycast Distribution due to UE Mobility

UEs' frequent moving from one 5G site to another can make it difficult to plan where the App Servers should be hosted. When one App server is heavily utilized, other App servers of the same address close-by can be very under-utilized. Since the condition can be short-lived, it is difficult for the application controller to anticipate the move and adjust.

1.4. Problem 3: Application Server Relocation

When an Application Server is added to, moved, or deleted from a 5G Edge Computing Data Center, not only the reachability changes but also the utilization and capacity for the Data Center might change.

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

2. Conventions used in this document

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

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

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

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

EC: Edge Computing

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

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

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

gNB next generation Node B

LDN: Local Data Network

PSA: PDU Session Anchor (UPF)

SSC: Session and Service Continuity

UE: User Equipment

UPF: User Plane Function

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

3. Solution Overview

From IP Layer, the Application Servers are identified by their IP (ANYCAST) addresses. To a router, having multiple servers with the same (ANYCAST) address attached to different egress routers (A-ER) is same as having multiple paths to reach the (ANYCAST) address.

There are many tools available to influence the path section on a router, such as the routing distance, TE metrics, policies, etc. This draft describes a solution to add "Site-Cost" to influence the path selection. The "Site-Cost", which is derived from "site-capacity + load measurement + Preference + xxx", can be raw measurements collected by the egress routers based on the instructions from a controller or can be informed by the App Controller periodically.

The proposed solution is for the egress router (A-ER) that have a direct connection to the Application Servers to collect desired measurements about the Servers' running status and advertise the metrics to other routers in 5G EC LDN.

The solution assumes that the 5G Edge Computing controller or management system is aware of the ANYCAST addresses that need optimized forwarding. To minimize the processing on routers, only the application flows that match with the ACLs configured by the 5G Edge Computing controller will collect and advertise the desired measurements.

3.1. Flow Affinity to an ANYCAST server

Having multiple Edge Computing Servers or App Layer Load Balancers with the same ANYCAST address attached to multiple A-ERs, Flow Affinity means routers sending the packets of the same flow to the same A-ER even if the cost towards the A-ER is no longer optimal.

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

Editor's note: for IPv6 traffic, Flow Affinity can be supported by the routers of the Local Data Network (LDN) forwarding the packets with the same Flow Label in the packets' IPv6 Header along the same path towards the same egress router.

3.2. IP Layer Metrics to Gauge App Server Running Status

Most applications do not expose their internal logic to the network. Their communications are generally encrypted. Most of them do not even respond to PING or ICMP messages initiated by routers or network gears.

[5G-EC-Metrics] describes the IP Layer Metrics that can gauge the application servers running status and environment:

- IP-Layer Metric for App Server Load Measurement:
  The Load Measurement to an App Server is a weighted combination of the number of packets/bytes to the App Server and the number of packets/bytes from the App Server which are collected by the A-ER that has the direct connection to the App Server.
  The A-ER is configured with an ACL that can filter out the packets for the Application Server.
- Capacity Index:
  Capacity Index is used to differentiate the running environment of the attached application server. Some data centers can have hundreds, or thousands, of servers behind an application server's App Layer Load Balancer. Other data centers can have a very small number of servers for the application. "Capacity

Index", which is a numeric number, is used to represent the capacity of the application server attached to an A-ER.

- Site preference index:
  [IPv6-StickyService] describes a scenario that some sites are more preferred for handling an application than others for flows from a specific UE.

For ease of description, those metrics, more may be added later, are called IP Layer App-Metrics throughout the document.

3.3. To Equalize traffic among Multiple ANYCAST Locations

The main benefit of using ANYCAST is to leverage the network layer information to balance the traffic among multiple Application Server locations.

For 5G Edge Computing environment, the routers in the LDN need to be notified of various measurements of the App Servers attached to each A-ER to make the intelligent decision on where to forward the traffic for the application from UEs.

[5G-EC-Metrics] describes the algorithms that can be used by the routers in LDN to compare the cost to reach the App Servers between the Site-i or Site-j:

```
               Load-i * CP-j               Pref-j * Network-Delay-i
Cost-i=min(w *(----------------) + (1-w) *(-------------------------))
              Load-j * CP-i              Pref-i * Network-Delay-j
```

Load-i: Load Index at Site-i, it is the weighted combination of the total packets or/and bytes sent to and received from the Application Server at Site-i during a fixed time period.

CP-i: capacity index at site I, a higher value means higher capacity.

Network Delay-i: Network latency measurement (RTT) to the A-ER that has the Application Server attached at the site-i.

Noted: Ingress nodes can easily measure RTT to all the egress nodes by existing IPPM metrics. But it is not so easy for ingress nodes to measure RTT to all the App Servers. Therefore, "Network-Delay-i", a.k.a. Network latency measurement (RTT), is between the Ingress nodes and egress nodes. The link cost between the egress nodes to their attached servers are embedded in the "capacity index".

Pref-i: Preference index for site-i, a higher value means higher preference.

w: Weight for load and site information, which is a value between 0 and 1. If smaller than 0.5, Network latency and the site Preference have more influence; otherwise, Server load and its capacity have more influence.

3.4. Reason for using IGP Based Solution

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

- Intermediate routers can derive the aggregated cost to reach the Application Servers attached to different egress nodes, especially:
  - The path to the optimal egress node can be more accurate or shorter
  - Convergence is shorter when there is any failure along the way towards the optimal ANYCAST server.
  - When there is any failure at the intended ANYCAST server, all the transient packets can be optimally forwarded to another App Server attached to a different egress router.
- Doesn't need the ingress nodes to establish tunnels with egress nodes.

There are limitations of using IGP too, such as:

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

4. Aggregated Cost Computed by Egress Routers

If all egress routers that have a direct connection to the App Servers can get a periodic update of the aggregated cost to the App Servers or can be configured with a consistent algorithm to compute an aggregated cost that takes into consideration the Load Measurement, Capacity value, and Preference value, this aggregated cost can be considered as the Metric of the link to the App Server.

In this scenario, there is no protocol extension needed.

4.1. OSPFv3 LSA to carry the Aggregated Cost

If the App Servers use IPv6 ANYCAST address, the aggregated cost computed by the egress routers can be encoded in the Metric field [the interface cost] of Intra-Area-Prefix-LSA specified by Section 3.7 of the [RFC5340].

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     6 (Intra-Area Prefix)     |         TLV Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          0    | Aggregated Cost to the App Server             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PrefixLength  | PrefixOptions |             0                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Address Prefix                      |
|                               ...                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Aggregated Cost to App Server

4.2. OSPFv2 LSA to carry the Aggregated Cost

For App Servers in IPv4 address, the Aggregated Cost can be encoded in the "Metric" field of the Stub Link LSA [Link type =3] specified by Section 12.4 of the [RFC2328].

5. IP Layer App-Metrics Advertisements

This section describes the OSPF extension that can carry the detailed IP Layer Metrics when it is not possible for all the egress routers to have a consistent algorithm to compute the aggregated cost or some routers need all the detailed IP Layer metrics for the App Servers for other purposes.

Since only a subset of routers within an IGP domain need to know those detailed metrics, it makes sense to use the OSPFv2 Extended Prefix Opaque LSA for IPv4 and OSPFv3 Extended LSA with Intra-Area-Prefix TLV to carry the detailed sub-TLVs. For routers that don't care about those metrics, they can ignore them very easily.

It worth noting that not all hosts (prefix) attached to an A-ER are ANYCAST servers that need network optimization. An A-ER only needs to advertise the App-Metrics for the ANYCAST addresses that match with the configured ACLs.

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

5.1. OSPFv3 Extension to carry the App-Metrics

For App Servers using IPv6, the OSPFv3 Extended LSA with the Intra-Area-Prefix Address TLV specified by the Section 3.7 of RFC8362 can be used to carry the App-Metrics for the attached App Servers.

```
  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
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |7 (IPv6 Local-Local Address)    |             Length            |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |            IPv6 AppServer (ANYCAST) address                   |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |            Load measurement sub-TLV                           |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |            Capability sub-TLV                                 |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |             Preference sub-TLV                                |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: IPv6 App Server App-Metrics Encoding

5.2. OSPFv2 Extension to advertise the IP Layer App-Metrics

For App Servers using IPv4 addresses, the OSPFv2 Extended Prefix Opaque LSA with the extended Prefix TLV can be used to carry the App Metrics sub-TLVs, as specified by the Section 2.1 [RFC7684].

Here is the proposed encoding:

```
  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                        |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Route Type    | Prefix Length | AF            | Flags         |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Address Prefix (variable)                                     |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Load Measurement Sub-TLV                                      |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | capacity Index Sub-TLV                                        |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Site Preference Sub-TLV                                       |
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: App-Metrix Sub-TLVs in OSPFv2 Extended Prefix TLV

5.3. IP Layer App-Metrics Sub-TLVs

Two types of Load Measurement Sub-TLVs are specified:

a) The Aggregated Load Index based on a weighted combination of the collected measurements;
b) The raw measurements of packets/bytes to/from the App Server address. The raw measurement is useful when the egress routers cannot be configured with a consistent algorithm to compute the aggregated load index or the raw measurements are needed by a central analytic system.

The Aggregated Load Index Sub-TLV has the following format:

```
 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 (TBD2)           |               Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Measurement Period                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Aggregated Load Index to reach the App Server       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Aggregated Load Index Sub-TLV

Type=TBD2 (to be assigned by IANA) indicates that the sub-TLV carries the Aggregated Load Measurement Index derived from the Weighted combination of bytes/packets sent to/received from the App server:

Index=w1*ToPackets+w2*FromPackes+w3*ToBytes+w4*FromBytes

Where wi is a value between 0 and 1; w1+ w2+ w3+ w4 = 1.

The Raw Load Measurement sub-TLV has the following format:

```
 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 (TBD3)         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Measurement Period                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of packets to the AppServer            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of packets from the AppServer          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of bytes to the AppServer              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of bytes from the AppServer            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      Figure 6: Raw Load Measurement Sub-TLV
```

Type= TBD3 (to be assigned by IANA) indicates that the sub-TLV carries the Raw measurements of packets/bytes to/from the App Server ANYCAST address.

Measurement Period: A user-specified period in seconds, default is 3600 seconds.

The Capacity Index sub-TLV has the following format:

```
  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 (TBD3)         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Capacity Index                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        Figure 7: Capacity Index Sub-TLV
```

The Preference Index sub-TLV has the following format:

```
 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 (TBD4)         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Preference Index                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: Preference Index Sub-TLV

Note: "Capacity Index" and "Site preference" can be more stable for each site. If those values are configured to nodes, they might not need to be included in every OSPF LSA.

6. Manageability Considerations

To be added.

7. Security Considerations

To be added.

8. IANA Considerations

The following Sub-TLV types need to be added by IANA to OSPFv4 Extended-LSA Sub-TLVs and OSPFv2 Extended Link Opaque LSA TLVs Registry.

- Aggregated Load Index Sub-TLV type
- Raw Load Measurement Sub-TLV type
- Capacity Index Sub-TLV type
- Preference Index Sub-TLV type

9. References

9.1. Normative References

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

[RFC2328] J. Moy, "OSPF Version 2", RFC 2328, April 1998.

[RFC7684] P. Psenak, et al, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, Nov. 2015.

[RFC8200] S. Deering R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", July 2017.

[RFC8326] A. Lindem, et al, "OSPFv3 Link State advertisement (LSA0 Extensibility", RFC 8362, April 2018.

9.2. Informative References

[3GPP-EdgeComputing] 3GPP TR 23.748, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on enhancement of support for Edge Computing in 5G Core network (5GC)", Release 17 work in progress, Aug 2020.

[5G-AppMetaData] L. Dunbar, K. Majumdar, H. Wang, "BGP NLRI App Meta Data for 5G Edge Computing Service", draft-dunbar-idr-5g-edge-compute-app-meta-data-01, work-in-progress, Nov 2020.

[5G-EC-Metrics] L. Dunbar, H. Song, J. Kaippallimalil, "IP Layer Metrics for 5G Edge Computing Service", draft-dunbar-ippm-5g-edge-compute-ip-layer-metrics-01, work-in-progress, Nov 2020.

[5G-StickyService] L. Dunbar, J. Kaippallimalil, "IPv6 Solution for 5G Edge Computing Sticky Service", draft-dunbar-6man-5g-ec-sticky-service-00, work-in-progress, Oct 2020.

[RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.

[BGP-SDWAN-Port] L. Dunbar, H. Wang, W. Hao, "BGP Extension for SDWAN Overlay Networks", draft-dunbar-idr-bgp-sdwan-overlay-ext-03, work-in-progress, Nov 2018.

[SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

10. Acknowledgments

Acknowledgements to Acee Lindem, Gyan Mishra, Jeff Tantsura, and Donald Eastlake for their review and suggestions.

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

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

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

Aijun Wang
China Telecom
Email: wangaj3@chinatelecom.cn

SPRING S. Hegde
Internet-Draft W. Britto
Intended status: Standards Track R. Shetty
Expires: October 18, 2021 Juniper Networks Inc.
B. Decraene
Orange
P. Psenak
Cisco Systems
T. Li
Arista Networks
April 16, 2021

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

Abstract

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

Requirements Language

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

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

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

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

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

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

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

2. Generic Metric Advertisement

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

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

2.1. ISIS Generic Metric sub-TLV

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

TLV-22 (Extended IS reachability) [RFC5305]

TLV-222 (MT-ISN) [RFC5120]

TLV-23 (IS Neighbor Attribute) [RFC5311]

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

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

sub-TLV 16 (Application-Specific Link Attributes) of TLV 22/222/23/223/141 [RFC8919]

```
 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     |  metric-type  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Value                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

   Type  :   TBD (To be assigned by IANA)
   Length: 4 octets
   metric-type:  A value from the IGP metric-type registry
   Value : metric value range (1 - 16,777,215)
```

Figure 1: ISIS Generic Metric sub-TLV

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

Default Metric, as of this writing), the Generic Metric advertisement MUST be ignored.

2.2. OSPF Generic Metric sub-TLV

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

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

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

sub-sub-TLV of Application-Specific Link Attributes sub-TLV [RFC 8920]

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

```
 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          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| metric-type   |         Reserved                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              Value                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Type  :   TBD (To be assigned by IANA)
Length: 8 octets
metric-type = A value from the IGP metric type registry

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

Figure 2: OSPF Generic Metric sub-TLV

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

3. FAD constraint sub-TLVs

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

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

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

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

3.1. ISIS FAD constraint sub-TLVs

3.1.1. ISIS Exclude Minimum Bandwidth sub-TLV

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

```
 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     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Min Bandwidth                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: TBA

Length: 4 octets.

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

Figure 3: ISIS FAEMB sub-TLV

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

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

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

3.1.2. ISIS Exclude Maximum Delay sub-TLV

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

```
 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     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      max link delay           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
where:
```

Type: TBD

Length: 3 octets

Max link delay: Maximum link delay in microseconds

Figure 4: ISIS FAEMD sub-TLV

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

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

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

3.2. OSPF FAD constraint sub-TLVs

3.2.1. OSPF Exclude Minimum Bandwidth sub-TLV

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

```
 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                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Min Bandwidth                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
where:
```

Type: TBD

Length: 4 octets.

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

Figure 5: OSPF FAEMB sub-TLV

The FAEMB sub-TLV MUST appear only once in the FAD sub-TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver. The Maximum Link Bandwidth as advertised by the sub-sub-TLV 23 of ASLA [RFC 8920] MUST be compared against the Minimum bandwidth advertised in FAEMB sub-TLV. If the link bandwidth is lower than the Minimum bandwidth advertised in FAEMB sub-TLV, the link MUST be excluded from the Flex-Algorithm topology. If a link does not have the Maximum Link Bandwidth advertised but the FAD contains this sub-TLV, then that link MUST be included in the topology and proceed to apply further pruning rules for the link.

3.2.2. OSPF Exclude Maximum Delay sub-TLV

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

```
 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                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Max Delay                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
where:

   Type: TBD

   Length:  3 octets

   Max link delay:  Maximum link delay in microseconds
```

Figure 6: OSPF FAEMD sub-TLV

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

4. Bandwidth Metric Advertisement

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

This document defines a new standard metric type for this purpose called the "Bandwidth Metric". The Bandwidth Metric MAY be advertised in the Generic Metric sub-TLV with the metric type set to "Bandwidth Metric". ISIS and OSPF will advertise this new type of metric in their link advertisements. Bandwidth metric is a link attribute and for advertisement and processing of this attribute for Flex-algorithm purposes, MUST follow the the section 12 of [I-D.ietf-lsr-flex-algo]

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

4.1. Automatic Metric Calculation

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

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

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

4.1.1. Automatic Metric Calculation Modes

4.1.1.1. Simple Mode

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

4.1.1.2. Interface Group Mode

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

```
A------B====C====F====D
       |              |
        ------E-------
```

Figure 7: Parallel interfaces

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

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

4.1.2. Automatic Metric Calculation Methods

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

1. Reference bandwidth method

2. Bandwidth thresholds method

4.1.2.1. Reference Bandwidth method

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

For example,

reference bandwidth = 1000G

Granularity = 20G

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

4.1.2.2. Bandwidth Thresholds method

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

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

4.1.3.1. Reference Bandwidth sub-TLV

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

```
 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     |     Flags     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Reference Bandwidth                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Granularity Bandwidth                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: TBD

Length: 9 octets.
Reference Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.
Granularity Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.

Flags:

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

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

```
Metric calculation: (Reference_bandwidth) /
                     (Total_link_bandwidth -
                     (Modulus of(Total_link_bandwidth,granularity_bw)))
```

Figure 8: ISIS FADRB sub-TLV

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

4.1.3.2. Bandwidth Thresholds sub-TLV

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

```
 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     |       Flags   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold 1                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric 1        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold 1                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric 2        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold 2                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                               .....
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric n-1      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold n-1                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric n        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: TBD

Length: 1 + n*7 octets. Here n is equal to number of Threshold Metrics spec
ified.
n MUST be greater than or equal to 1.

Flags:

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

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

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

Figure 9: ISIS FADBT sub-TLV

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

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

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

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

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

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

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

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

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

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

4.1.4.1. Reference Bandwidth sub-TLV

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

```
 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                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Flags      |      Reserved                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Reference Bandwidth                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Granularity Bandwidth                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: TBD

Length: 14 octets.
Reference Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.
Granularity Bandwidth: Bandwidth encoded in 32 bits in IEEE floating point format. The units are in bytes per second.

Flags:

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

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

```
Metric calculation: (Reference_bandwidth) /
                     (Total_link_bandwidth -
                     (Modulus of(Total_link_bandwidth, Granularity_bw)))
```

Figure 10: OSPF FADRB sub-TLV

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

4.1.4.2. Bandwidth Threshold sub-TLV

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

```
 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                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Flags     | Reserved                                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold 1                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric 1                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold 2                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric 2                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold 3                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                              .....
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric n-1                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Bandwidth Threshold n                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Threshold Metric n                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: TBD

Length: 2 + n*8 octets. Here n is equal to number of Threshold Metrics specified.
n MUST be greater than or equal to 1.

Flags:

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

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

Staircase bandwidth threshold and associated metric values.
Bandwidth Threshold 1: Minimum Link Bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.
Bandwidth Threshold 2: Maximum Link Bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.
Threshold Metric 1 : metric value range (1 - 4,294,967,296)

Figure 11: OSPF FADBT sub-TLV

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

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

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

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

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

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

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

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

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

5. Bandwidth metric considerations

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

1. If the the Generic Metric sub-TLV with Bandwidth metric type is advertised for the link as described in Section 4, it MUST be used during the Flex-Algorithm calculation.

2. If the Generic Metric sub-TLV with Bandwidth metric type is not advertised for the link and the winning FAD for the Flex-Algorithm does not specify the automatic bandwidth metric calculation (as defined in Section 4.1), the Bandwidth Metric is considered as not being advertised for the link.

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

6. Calculation of Flex-Algorithm paths

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

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

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

7. Backward Compatibility

8. Security Considerations

TBD

9. IANA Considerations

9.1. IGP Metric-Type Registry

Type: Suggested 3 (TBA)

Description: Bandwidth metric

Reference: This document

Type: 128 to 255(TBA)

Description: User defned metric

Reference: This document

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

Type: Suggested 6 (TBA)

Description: ISIS Exclude Minimum Bandwidth sub-TLV

Reference: This document Section 3.1.1

Type: Suggested 7 (TBA)

Description: ISIS Exclude Maximum Delay sub-TLV

Reference: This document Section 3.1.2

Type: Suggested 8 (TBA)

Description: ISIS Reference Bandwidth sub-TLV

Reference: This document Section 4.1.3.1

Type: Suggested 9 (TBA)

Description: ISIS Threshold Metric sub-TLV

Reference: This document Section 4.1.3.2

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

Type: Suggested 6 (TBA)

Description: OSPF Exclude Minimum Bandwidth sub-TLV

Reference: This document Section 3.2.1

Type: Suggested 7 (TBA)

Description: OSPF Exclude Maximum Delay sub-TLV

Reference: This document Section 3.2.2

Type: Suggested 8 (TBA)

Description: OSPF Reference Bandwidth sub-TLV

Reference: This document Section 4.1.4.1

Type: Suggested 9 (TBA)

Description: OSPF Threshold Metric sub-TLV

Reference: This document Section 4.1.4.2

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

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.1

9.5. Sub-sub-TLV Codepoints for Application-Specific Link Attributes

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.1

9.6. OSPFv2 Extended Link TLV Sub-TLVs

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.2

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

Type: Suggested 45 (TBA)

Description: Generic metric

Reference: This document Section 2.2

10. Acknowledgements

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

11. Contributors

1. Salih K A

Juniper Networks

salih@juniper.net

12. References

12.1. Normative References

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

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

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

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

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

12.2. Informative References

[I-D.bashandy-rtgwg-segment-routing-uloop]
Bashandy, A., Filsfils, C., Litkowski, S., Decraene, B., Francois, P., and P. Psenak, "Loop avoidance using Segment Routing", draft-bashandy-rtgwg-segment-routing-uloop-10 (work in progress), December 2020.

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

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

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

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

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

Authors' Addresses

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: shraddha@juniper.net

William Britto A J
Juniper Networks Inc.

Email: bwilliam@juniper.net

Rajesh Shetty
Juniper Networks Inc.

Email: mrajesh@juniper.net

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Peter Psenak
Cisco Systems

Email: ppsenak@cisco.com

Tony Li
Arista Networks

Email: tony.li@tony.li

Network Working Group P. Psenak, Ed.
Internet-Draft Cisco Systems
Intended status: Standards Track S. Hegde
Expires: October 30, 2021 Juniper Networks, Inc.
C. Filsfils
K. Talaulikar
Cisco Systems, Inc.
A. Gulko
Edward Jones
April 28, 2021

IGP Flexible Algorithm
draft-ietf-lsr-flex-algo-15

Abstract

IGP protocols traditionally compute best paths over the network based on the IGP metric assigned to the links. Many network deployments use RSVP-TE based or Segment Routing based Traffic Engineering to steer traffic over a path that is computed using different metrics or constraints than the shortest IGP path. This document proposes a solution that allows IGPs themselves to compute constraint-based paths over the network. This document also specifies a way of using Segment Routing (SR) Prefix-SIDs and SRv6 locators to steer packets along the constraint-based paths.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

An IGP-computed path based on the shortest IGP metric is often be replaced by a traffic-engineered path due to the traffic requirements which are not reflected by the IGP metric. Some networks engineer the IGP metric assignments in a way that the IGP metric reflects the link bandwidth or delay. If, for example, the IGP metric is reflecting the bandwidth on the link and the application traffic is delay sensitive, the best IGP path may not reflect the best path from such an application's perspective.

To overcome this limitation, various sorts of traffic engineering have been deployed, including RSVP-TE and SR-TE, in which case the TE component is responsible for computing paths based on additional metrics and/or constraints. Such paths need to be installed in the forwarding tables in addition to, or as a replacement for, the original paths computed by IGPs. Tunnels are often used to represent the engineered paths and mechanisms like one described in [RFC3906] are used to replace the native IGP paths with such tunnel paths.

This document specifies a set of extensions to ISIS, OSPFv2, and OSPFv3 that enable a router to advertise TLVs that identify (a) calculation-type, (b) specify a metric-type, and (c) describe a set of constraints on the topology, that are to be used to compute the best paths along the constrained topology. A given combination of calculation-type, metric-type, and constraints is known as a "Flexible Algorithm Definition". A router that sends such a set of TLVs also assigns a Flex-Algorithm value to the specified combination of calculation-type, metric-type, and constraints.

This document also specifies a way for a router to use IGPs to associate one or more SR Prefix-SIDs or SRv6 locators with a particular Flex-Algorithm. Each such Prefix-SID or SRv6 locator then represents a path that is computed according to the identified Flex-Algorithm.

2. Requirements Language

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

3. Terminology

This section defines terms that are often used in this document.

Flexible Algorithm Definition (FAD) - the set consisting of (a) calculation-type, (b) metric-type, and (c) a set of constraints.

Flexible Algorithm - a numeric identifier in the range 128-255 that is associated via configuration with the Flexible-Algorithm Definition.

Local Flexible Algorithm Definition - Flexible Algorithm Definition defined locally on the node.

Remote Flexible Algorithm Definition - Flexible Algorithm Definition received from other nodes via IGP flooding.

Flexible Algorithm Participation - per application configuration state that expresses whether the node is participating in a particular Flexible Algorithm.

IGP Algorithm - value from the the "IGP Algorithm Types" registry defined under "Interior Gateway Protocol (IGP) Parameters" IANA registries. IGP Algorithms represents the triplet (Calculation Type, Metric, Constraints), where the second and third elements of the triple MAY be unspecified.

ABR - Area Border Router. In ISIS terminology it is also known as L1/L2 router.

ASBR - Autonomous System Border Router.

4. Flexible Algorithm

Many possible constraints may be used to compute a path over a network. Some networks are deployed as multiple planes. A simple form of constraint may be to use a particular plane. A more sophisticated form of constraint can include some extended metric as described in [RFC8570]. Constraints which restrict paths to links with specific affinities or avoid links with specific affinities are also possible. Combinations of these are also possible.

To provide maximum flexibility, we want to provide a mechanism that allows a router to (a) identify a particular calculation-type, (b) metric-type, (c) describe a particular set of constraints, and (d) assign a numeric identifier, referred to as Flex-Algorithm, to the combination of that calculation-type, metric-type, and those constraints. We want the mapping between the Flex-Algorithm and its meaning to be flexible and defined by the user. As long as all routers in the domain have a common understanding as to what a particular Flex-Algorithm represents, the resulting routing computation is consistent and traffic is not subject to any looping.

The set consisting of (a) calculation-type, (b) metric-type, and (c) a set of constraints is referred to as a Flexible-Algorithm Definition.

Flexible-Algorithm is a numeric identifier in the range 128-255 that is associated via configuratin with the Flexible-Algorithm Definition.

IANA "IGP Algorithm Types" registry defines the set of values for IGP Algorithms. We propose to allocate the following values for Flex-Algorithms from this registry:

128-255 - Flex-Algorithms

5. Flexible Algorithm Definition Advertisement

To guarantee the loop-free forwarding for paths computed for a particular Flex-Algorithm, all routers that (a) are configured to participate in a particular Flex-Algorithm, and (b) are in the same Flex-Algorithm definition advertisement scope MUST agree on the definition of the Flex-Algorithm.

5.1. ISIS Flexible Algorithm Definition Sub-TLV

The ISIS Flexible Algorithm Definition Sub-TLV (FAD Sub-TLV) is used to advertise the definition of the Flex-Algorithm.

The ISIS FAD Sub-TLV is advertised as a Sub-TLV of the ISIS Router Capability TLV-242 that is defined in [RFC7981].

ISIS FAD Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Length     |Flex-Algorithm | Metric-Type   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Calc-Type   |    Priority   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Sub-TLVs                             |
+                                                               +
|                            ...                                |

|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 26

Length: variable, dependent on the included Sub-TLVs

Flex-Algorithm: Single octet value between 128 and 255 inclusive.

Metric-Type: Type of metric to be used during the calculation. Following values are defined:

0: IGP Metric

1: Min Unidirectional Link Delay as defined in [RFC8570], section 4.2, encoded as application specific link attribute as specified in [RFC8919] and Section 12 of this document.

2: Traffic Engineering Default Metric as defined in [RFC5305], section 3.7, encoded as application specific link attribute as specified in [RFC8919] and Section 12 of this document.

Calc-Type: value from 0 to 127 inclusive from the "IGP Algorithm Types" registry defined under "Interior Gateway Protocol (IGP) Parameters" IANA registries. IGP algorithms in the range of 0-127 have a defined triplet (Calculation Type, Metric, Constraints). When used to specify the Calc-Type in the FAD Sub-TLV, only the Calculation Type defined for the specified IGP Algorithm is used. The Metric/Constraints MUST NOT be inherited. If the required calculation type is Shortest Path First, the value 0 SHOULD appear in this field.

Priority: Value between 0 and 255 inclusive that specifies the priority of the advertisement.

Sub-TLVs - optional sub-TLVs.

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

The ISIS FAD Sub-TLV has an area scope. The Router Capability TLV in which the FAD Sub-TLV is present MUST have the S-bit clear.

ISIS L1/L2 router MAY be configured to re-generate the winning FAD from level 2, without any modification to it, to level 1 area. The re-generation of the FAD Sub-TLV from level 2 to level 1 is determined by the L1/L2 router, not by the originator of the FAD advertisement in the level 2. In such case, the re-generated FAD Sub-TLV will be advertised in the level 1 Router Capability TLV originated by the L1/L2 router.

L1/L2 router MUST NOT re-generate any FAD Sub-TLV from level 1 to level 2.

5.2. OSPF Flexible Algorithm Definition TLV

OSPF FAD TLV is advertised as a top-level TLV of the RI LSA that is defined in [RFC7770].

OSPF FAD TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Flex-Algorithm |  Metric-Type  |   Calc-Type   |    Priority   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Sub-TLVs                           |
+                                                               +
|                               ...                             |

|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 16

Length: variable, dependent on the included Sub-TLVs

Flex-Algorithm:: Flex-Algorithm number. Value between 128 and 255 inclusive.

Metric-Type: Type of metric to be used during the calculation. Following values are defined:

0: IGP Metric

1: Min Unidirectional Link Delay as defined in [RFC7471], section 4.2, encoded as application specific link attribute as specified in [RFC8920] and Section 12 of this document.

2: Traffic Engineering metric as defined in [RFC3630], section 2.5.5, encoded as application specific link attribute as specified in [RFC8920] and Section 12 of this document.

Calc-Type: as described in Section 5.1

Priority: as described in Section 5.1

Sub-TLVs - optional sub-TLVs.

When multiple OSPF FAD TLVs, for the same Flexible-Algorithm, are received from a given router, the receiver MUST use the first occurrence of the TLV in the Router Information LSA. If the OSPF FAD TLV, for the same Flex-Algorithm, appears in multiple Router Information LSAs that have different flooding scopes, the OSPF FAD TLV in the Router Information LSA with the area-scoped flooding scope MUST be used. If the OSPF FAD TLV, for the same algorithm, appears in multiple Router Information LSAs that have the same flooding scope, the OSPF FAD TLV in the Router Information (RI) LSA with the numerically smallest Instance ID MUST be used and subsequent instances of the OSPF FAD TLV MUST be ignored.

The RI LSA can be advertised at any of the defined opaque flooding scopes (link, area, or Autonomous System (AS)). For the purpose of OSPF FAD TLV advertisement, area-scoped flooding is REQUIRED. The Autonomous System flooding scope SHOULD NOT be used by default unless local configuration policy on the originating router indicates domain wide flooding.

5.3. Common Handling of Flexible Algorithm Definition TLV

This section describes the protocol-independent handling of the FAD TLV (OSPF) or FAD Sub-TLV (ISIS). We will refer to it as FAD TLV in this section, even though in case of ISIS it is a Sub-TLV.

The value of the Flex-Algorithm MUST be between 128 and 255 inclusive. If it is not, the FAD TLV MUST be ignored.

Only a subset of the routers participating in the particular Flex-Algorithm need to advertise the definition of the Flex-Algorithm.

Every router, that is configured to participate in a particular Flex-Algorithm, MUST select the Flex-Algorithm definition based on the following ordered rules. This allows for the consistent Flex-Algorithm definition selection in cases where different routers advertise different definitions for a given Flex-Algorithm:

1. From the advertisements of the FAD in the area (including both locally generated advertisements and received advertisements) select the one(s) with the highest priority value.

2. If there are multiple advertisements of the FAD with the same highest priority, select the one that is originated from the router with the highest System-ID, in the case of ISIS, or Router

ID, in the case of OSPFv2 and OSPFv3. For ISIS, the System-ID is described in [ISO10589]. For OSPFv2 and OSPFv3, standard Router ID is described in [RFC2328] and [RFC5340] respectively.

A router that is not configured to participate in a particular Flex-Algorithm MUST ignore FAD Sub-TLVs advertisements for such Flex-Algorithm.

A router that is not participating in a particular Flex-Algorithm is allowed to advertise FAD for such Flex-Algorithm. Receiving routers MUST consider FAD advertisement regardless of the Flex-Algorithm participation of the FAD originator.

Any change in the Flex-Algorithm definition may result in temporary disruption of traffic that is forwarded based on such Flex-Algorithm paths. The impact is similar to any other event that requires network-wide convergence.

If a node is configured to participate in a particular Flexible-Algorithm, but the selected Flex-Algorithm definition includes calculation-type, metric-type, constraint, flag, or Sub-TLV that is not supported by the node, it MUST stop participating in such Flexible-Algorithm. That implies that it MUST NOT announce participation for such Flexible-Algorithm as specified in Section 11 and it MUST remove any forwarding state associated with it.

Flex-Algorithm definition is topology independent. It applies to all topologies that a router participates in.

6. Sub-TLVs of ISIS FAD Sub-TLV

6.1. ISIS Flexible Algorithm Exclude Admin Group Sub-TLV

The Flexible Algorithm definition can specify 'colors' that are used by the operator to exclude links during the Flex-Algorithm path computation.

The ISIS Flexible Algorithm Exclude Admin Group Sub-TLV is used to advertise the exclude rule that is used during the Flex-Algorithm path calculation as specified in Section 13.

The ISIS Flexible Algorithm Exclude Admin Group Sub-TLV (FAEAG Sub-TLV) is a Sub-TLV of the ISIS FAD Sub-TLV. It has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Length     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Extended Admin Group                     |
+-                                                             -+
|                              ...                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 1

Length: variable, dependent on the size of the Extended Admin Group. MUST be a multiple of 4 octets.

Extended Administrative Group: Extended Administrative Group as defined in [RFC7308].

The ISIS FAEAG Sub-TLV MUST NOT appear more than once in an ISIS FAD Sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

6.2. ISIS Flexible Algorithm Include-Any Admin Group Sub-TLV

The Flexible Algorithm definition can specify 'colors' that are used by the operator to include links during the Flex-Algorithm path computation.

The ISIS Flexible Algorithm Include-Any Admin Group Sub-TLV is used to advertise include-any rule that is used during the Flex-Algorithm path calculation as specified in Section 13.

The format of the ISIS Flexible Algorithm Include-Any Admin Group Sub-TLV is identical to the format of the FAEAG Sub-TLV in Section 6.1.

The ISIS Flexible Algorithm Include-Any Admin Group Sub-TLV Type is 2.

The ISIS Flexible Algorithm Include-Any Admin Group Sub-TLV MUST NOT appear more than once in an ISIS FAD Sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

6.3. ISIS Flexible Algorithm Include-All Admin Group Sub-TLV

The Flexible Algorithm definition can specify 'colors' that are used by the operator to include link during the Flex-Algorithm path computation.

The ISIS Flexible Algorithm Include-All Admin Group Sub-TLV is used to advertise include-all rule that is used during the Flex-Algorithm path calculation as specified in Section 13.

The format of the ISIS Flexible Algorithm Include-All Admin Group Sub-TLV is identical to the format of the FAEAG Sub-TLV in Section 6.1.

The ISIS Flexible Algorithm Include-All Admin Group Sub-TLV Type is 3.

The ISIS Flexible Algorithm Include-All Admin Group Sub-TLV MUST NOT appear more than once in an ISIS FAD Sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

6.4. ISIS Flexible Algorithm Definition Flags Sub-TLV

The ISIS Flexible Algorithm Definition Flags Sub-TLV (FADF Sub-TLV) is a Sub-TLV of the ISIS FAD Sub-TLV. It has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Length     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Flags                             |
+-                                                             -+
|                            ...                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 4

Length: variable, non-zero number of octets of the Flags field

Flags:

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

M-flag: when set, the Flex-Algorithm specific prefix metric MUST be used for inter-area and external prefix calculation. This flag is not applicable to prefixes advertised as SRv6 locators.

Bits are defined/sent starting with Bit 0 defined above. Additional bit definitions that may be defined in the future SHOULD be assigned in ascending bit order so as to minimize the number of bits that will need to be transmitted.

Undefined bits MUST be transmitted as 0.

Bits that are NOT transmitted MUST be treated as if they are set to 0 on receipt.

The ISIS FADF Sub-TLV MUST NOT appear more than once in an ISIS FAD Sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

If the ISIS FADF Sub-TLV is not present inside the ISIS FAD Sub-TLV, all the bits are assumed to be set to 0.

6.5. ISIS Flexible Algorithm Exclude SRLG Sub-TLV

The Flexible Algorithm definition can specify Shared Risk Link Groups (SRLGs) that the operator wants to exclude during the Flex-Algorithm path computation.

The ISIS Flexible Algorithm Exclude SRLG Sub-TLV (FAESRLG) is used to advertise the exclude rule that is used during the Flex-Algorithm path calculation as specified in Section 13.

The ISIS FAESRLG Sub-TLV is a Sub-TLV of the ISIS FAD Sub-TLV. It has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Length     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Shared Risk Link Group Value             |
+-                                                             -+
|                            ...                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
where:

Type: 5

Length: variable, dependent on number of SRLG values. MUST be a multiple of 4 octets.

Shared Risk Link Group Value: SRLG value as defined in [RFC5307].

The ISIS FAESRLG Sub-TLV MUST NOT appear more than once in an ISIS FAD Sub-TLV. If it appears more than once, the ISIS FAD Sub-TLV MUST be ignored by the receiver.

7. Sub-TLVs of OSPF FAD TLV

7.1. OSPF Flexible Algorithm Exclude Admin Group Sub-TLV

The Flexible Algorithm Exclude Admin Group Sub-TLV (FAEAG Sub-TLV) is a Sub-TLV of the OSPF FAD TLV. It's usage is described in Section 6.1. It has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Extended Admin Group                     |
+-                                                             -+
|                            ...                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 1

Length: variable, dependent on the size of the Extended Admin Group. MUST be a multiple of 4 octets.

Extended Administrative Group: Extended Administrative Group as defined in [RFC7308].

The OSPF FAEAG Sub-TLV MUST NOT appear more than once in an OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver.

7.2. OSPF Flexible Algorithm Include-Any Admin Group Sub-TLV

The usage of this Sub-TLVs is described in Section 6.2.

The format of the OSPF Flexible Algorithm Include-Any Admin Group Sub-TLV is identical to the format of the OSPF FAEAG Sub-TLV in Section 7.1.

The OSPF Flexible Algorithm Include-Any Admin Group Sub-TLV Type is 2.

The OSPF Flexible Algorithm Include-Any Admin Group Sub-TLV MUST NOT appear more than once in an OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver.

7.3. OSPF Flexible Algorithm Include-All Admin Group Sub-TLV

The usage of this Sub-TLVs is described in Section 6.3.

The format of the OSPF Flexible Algorithm Include-Any Admin Group Sub-TLV is identical to the format of the OSPF FAEAG Sub-TLV in Section 7.1.

The OSPF Flexible Algorithm Include-Any Admin Group Sub-TLV Type is 3.

The OSPF Flexible Algorithm Include-All Admin Group Sub-TLV MUST NOT appear more than once in an OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver.

7.4. OSPF Flexible Algorithm Definition Flags Sub-TLV

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Flags                             |
+-                                                             -+
|                            ...                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 4

Length: variable, dependent on the size of the Flags field. MUST be a multiple of 4 octets.

Flags:

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

M-flag: when set, the Flex-Algorithm specific prefix and ASBR metric MUST be used for inter-area and external prefix calculation. This flag is not applicable to prefixes advertised as SRv6 locators.

Bits are defined/sent starting with Bit 0 defined above. Additional bit definitions that may be defined in the future SHOULD be assigned in ascending bit order so as to minimize the number of bits that will need to be transmitted.

Undefined bits MUST be transmitted as 0.

Bits that are NOT transmitted MUST be treated as if they are set to 0 on receipt.

The OSPF FADF Sub-TLV MUST NOT appear more than once in an OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver.

If the OSPF FADF Sub-TLV is not present inside the OSPF FAD TLV, all the bits are assumed to be set to 0.

7.5. OSPF Flexible Algorithm Exclude SRLG Sub-TLV

The OSPF Flexible Algorithm Exclude SRLG Sub-TLV (FAESRLG Sub-TLV) is a Sub-TLV of the OSPF FAD TLV. Its usage is described in Section 6.5. It has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Shared Risk Link Group Value                  |
+-                                                             -+
|                            ...                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
where:

Type: 5

Length: variable, dependent on the number of SRLGs. MUST be a multiple of 4 octets.

Shared Risk Link Group Value: SRLG value as defined in [RFC4203].

The OSPF FAESRLG Sub-TLV MUST NOT appear more than once in an OSPF FAD TLV. If it appears more than once, the OSPF FAD TLV MUST be ignored by the receiver.

8. ISIS Flexible Algorithm Prefix Metric Sub-TLV

The ISIS Flexible Algorithm Prefix Metric (FAPM) Sub-TLV supports the advertisement of a Flex-Algorithm specific prefix metric associated with a given prefix advertisement.

The ISIS FAPM Sub-TLV is a sub-TLV of TLVs 135, 235, 236, and 237 and has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Length     |Flex-Algorithm |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Metric                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 6

Length: 5 octets

Flex-Algorithm: Single octet value between 128 and 255 inclusive.

Metric: 4 octets of metric information

The ISIS FAPM Sub-TLV MAY appear multiple times in its parent TLV. If it appears more than once with the same Flex-Algorithm value, the first instance MUST be used and any subsequent instances MUST be ignored.

If a prefix is advertised with a Flex-Algorithm prefix metric larger then MAX_PATH_METRIC as defined in [RFC5305] this prefix MUST NOT be considered during the Flexible-Algorithm computation.

The usage of the Flex-Algorithm prefix metric is described in Section 13.

The ISIS FAPM Sub-TLV MUST NOT be advertised as a sub-TLV of the ISIS SRv6 Locator TLV [I-D.ietf-lsr-isis-srv6-extensions]. The ISIS SRv6 Locator TLV includes the Algorithm and Metric fields which MUST be used instead. If the FAPM Sub-TLV is present as a sub-TLV of the

ISIS SRv6 Locator TLV in the received LSP, such FAPM Sub-TLV MUST be ignored.

9. OSPF Flexible Algorithm Prefix Metric Sub-TLV

The OSPF Flexible Algorithm Prefix Metric (FAPM) Sub-TLV supports the advertisement of a Flex-Algorithm specific prefix metric associated with a given prefix advertisement.

The OSPF Flex-Algorithm Prefix Metric (FAPM) Sub-TLV is a Sub-TLV of the:

- OSPFv2 Extended Prefix TLV [RFC7684]

- Following OSPFv3 TLVs as defined in [RFC8362]:

  Inter-Area Prefix TLV

  External Prefix TLV

OSPF FAPM Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Flex-Algorithm |     Flags     |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Metric                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 3 for OSPFv2, 26 for OSPFv3

Length: 8 octets

Flex-Algorithm: Single octet value between 128 and 255 inclusive.

Flags: single octet value

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

E bit : position 0: The type of external metric. If bit is set, the metric specified is a Type 2 external metric. This bit is applicable only to OSPF External and NSSA external prefixes. This is semantically the same as E bit in section A.4.5 of [RFC2328] and section A.4.7 of [RFC5340] for OSPFv2 and OSPFv3 respectively.

Bits 1 through 7: MUST be cleared by sender and ignored by receiver.

Reserved: Must be set to 0, ignored at reception.

Metric: 4 octets of metric information

The OSPF FAPM Sub-TLV MAY appear multiple times in its parent TLV. If it appears more than once with the same Flex-Algorithm value, the first instance MUST be used and any subsequent instances MUST be ignored.

The usage of the Flex-Algorithm prefix metric is described in Section 13.

10. OSPF Flexible Algorithm ASBR Reachability Advertisement

An OSPF ABR advertises the reachability of ASBRs in its attached areas to enable routers within those areas to perform route calculations for external prefixes advertised by the ASBRs. OSPF extensions for advertisement of Flex-Algorithm specific reachability and metric for ASBRs is similarly required for Flex-Algorithm external prefix computations as described further in Section 13.1.

10.1. OSPFv2 Extended Inter-Area ASBR LSA

The OSPFv2 Extended Inter-Area ASBR (EIA-ASBR) LSA is an OSPF Opaque LSA [RFC5250] that is used to advertise additional attributes related to the reachability of the OSPFv2 ASBR that is external to the area yet internal to the OSPF domain. Semantically, the OSPFv2 EIA-ASBR LSA is equivalent to the fixed format Type 4 Summary LSA [RFC2328]. Unlike the Type 4 Summary LSA, the LSID of the EIA-ASBR LSA does not carry the ASBR Router-ID - the ASBR Router-ID is carried in the body of the LSA. OSPFv2 EIA-ASBR LSA is advertised by an OSPFv2 ABR and its flooding is defined to be area-scoped only.

An OSPFv2 ABR generates the EIA-ASBR LSA for an ASBR when it is advertising the Type-4 Summary LSA for it and has the need for advertising additional attributes for that ASBR beyond what is conveyed in the fixed format Type-4 Summary LSA. An OSPFv2 ABR MUST NOT advertise the EIA-ASBR LSA for an ASBR for which it is not advertising the Type 4 Summary LSA. This ensures that the ABR does not generate the EIA-ASBR LSA for an ASBR to which it does not have reachability in the base OSPFv2 topology calculation. The OSPFv2 ABR SHOULD NOT advertise the EIA-ASBR LSA for an ASBR when it does not have additional attributes to advertise for that ASBR.

The OSPFv2 EIA-ASBR LSA has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            LS age             |     Options   |    LS Type    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Opaque Type  |                 Opaque ID                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Advertising Router                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     LS sequence number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         LS checksum           |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+-                            TLVs                             -+
|                             ...                               |
```

The Opaque Type used by the OSPFv2 EIA-ASBR LSA is TBD (suggested value 11). The Opaque Type is used to differentiate the various types of OSPFv2 Opaque LSAs and is described in Section 3 of [RFC5250]. The LS Type MUST be 10, indicating that the Opaque LSA flooding scope is area-local [RFC5250]. The LSA Length field [RFC2328] represents the total length (in octets) of the Opaque LSA, including the LSA header and all TLVs (including padding).

The Opaque ID field is an arbitrary value used to maintain multiple OSPFv2 EIA-ASBR LSAs. For OSPFv2 EIA-ASBR LSAs, the Opaque ID has no semantic significance other than to differentiate OSPFv2 EIA-ASBR LSAs originated by the same OSPFv2 ABR. If multiple OSPFv2 EIA-ASBR LSAs specify the same ASBR, the attributes from the Opaque LSA with the lowest Opaque ID SHOULD be used.

The format of the TLVs within the body of the OSPFv2 EIA-ASBR LSA is the same as the format used by the Traffic Engineering Extensions to OSPFv2 [RFC3630]. The variable TLV section consists of one or more nested TLV tuples. Nested TLVs are also referred to as sub- TLVs. The Length field defines the length of the value portion in octets (thus, a TLV with no value portion would have a length of 0). The TLV is padded to 4-octet alignment; padding is not included in the

Length field (so a 3-octet value would have a length of 3, but the total size of the TLV would be 8 octets). Nested TLVs are also 32-bit aligned. For example, a 1-byte value would have the Length field set to 1, and 3 octets of padding would be added to the end of the value portion of the TLV. The padding is composed of zeros.

10.1.1. OSPFv2 Extended Inter-Area ASBR TLV

The OSPFv2 Extended Inter-Area ASBR (EIA-ASBR) TLV is a top-level TLV of the OSPFv2 EIA-ASBR LSA and is used to advertise additional attributes associated with the reachability of an ASBR.

The OSPFv2 EIA-ASBR TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        ASBR Router ID                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
.                            Sub-TLVs                           .
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 1

Length: variable

ASBR Router ID: four octets carrying the OSPF Router ID of the ASBR whose information is being carried.

Sub-TLVs : variable

Only a single OSPFv2 EIA-ASBR TLV MUST be advertised in each OSPFv2 EIA-ASBR LSA and the receiver MUST ignore all instances of this TLV other than the first one in an LSA.

OSPFv2 EIA-ASBR TLV MUST be present inside an OSPFv2 EIA-ASBR LSA with at least a single sub-TLV included, otherwise the OSPFv2 EIA-ASBR LSA MUST be ignored by the receiver.

10.2. OSPF Flexible Algorithm ASBR Metric Sub-TLV

The OSPF Flexible Algorithm ASBR Metric (FAAM) Sub-TLV supports the advertisement of a Flex-Algorithm specific metric associated with a given ASBR reachability advertisement by an ABR.

The OSPF Flex-Algorithm ASBR Metric (FAAM) Sub-TLV is a Sub-TLV of the:

- OSPFv2 Extended Inter-Area ASBR TLV as defined in Section 10.1.1

- OSPFv3 Inter-Area-Router TLV defined in [RFC8362]

OSPF FAAM Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Type            |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Flex-Algorithm |                   Reserved                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Metric                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: 1 for OSPFv2, TBD (suggested value 30) for OSPFv3

Length: 8 octets

Flex-Algorithm: Single octet value between 128 and 255 inclusive.

Reserved: Must be set to 0, ignored at reception.

Metric: 4 octets of metric information

The OSPF FAAM Sub-TLV MAY appear multiple times in its parent TLV. If it appears more than once with the same Flex-Algorithm value, the first instance MUST be used and any subsequent instances MUST be ignored.

The advertisement of the ASBR reachability using the OSPF FAAM Sub-TLV inside the OSPFv2 EIA-ASBR LSA follows the section 12.4.3 of [RFC2328] and inside the OSPFv3 E-Inter-Area-Router LSA follows the section 4.8.5 of [RFC5340]. The reachability of the ASBR is evaluated in the context of the specific Flex-Algorithm.

The FAAM computed by the ABR will be equal to the metric to reach the ASBR for a given Flex-Algorithm in a source area or the cumulative metric via other ABR(s) when the ASBR is in a remote area. This is similar in nature to how the metric is set when the ASBR reachability metric is computed in the default algorithm for the metric in the OSPFv2 Type 4 ASBR Summary LSA and the OSPFv3 Inter-Area-Router LSA.

An OSPF ABR MUST NOT include the OSPF FAAM Sub-TLV with a specific Flex-Algorithm in its reachability advertisement for an ASBR between areas unless that ASBR is reachable for it in the context of that specific Flex-Algorithm.

An OSPF ABR MUST include the OSPF FAAM Sub-TLVs as part of the ASBR reachability advertisement between areas for the Flex-Algorithm for which the winning FAD includes the M-flag and the ASBR is reachable in the context of that specific Flex-Algorithm.

OSPF routers MUST use the OSPF FAAM Sub-TLV to calculate the reachability of the ASBRs if the winning FAD for the specific Flex-Algorithm includes the M-flag. OSPF routers MUST NOT use the OSPF FAAM Sub-TLV to calculate the reachability of the ASBRs for the specific Flex-Algorithm if the winning FAD for such Flex-Algorithm does not include the M-flag. Instead, the OSPFv2 Type 4 Summary LSAs or the OSPFv3 Inter-Area-Router-LSAs MUST be used instead as specified in section 16.2 of [RFC2328] and section 4.8.5 of [RFC5340] for OSPFv2 and OSPFv3 respectively.

The processing of the new or changed OSPF FAAM Sub-TLV triggers the processing of the External routes similar to what is described in section 16.5 of the [RFC2328] for OSPFv2 and section 4.8.5 of [RFC5340] for OSPFv3 for the specific Flex-Algorithm. The External and NSSA External route calculation should be limited to Flex-Algorithm(s) for which the winning FAD(s) includes the M-flag.

Processing of the OSPF FAAM Sub-TLV does not require the existence of the equivalent OSPFv2 Type 4 Summary LSA or the OSPFv3 Inter-Area-Router-LSA that is advertised by the same ABR inside the area. When the OSPFv2 EIA-ASBR LSA or the OSPFv3 E-Inter-Area-Router-LSA are advertised along with the OSPF FAAM Sub-TLV by the ABR for a specific ASBR, it is expected that the same ABR would advertise the reachability of the same ASBR in the equivalent base LSAs - i.e., the OSPFv2 Type 4 Summary LSA or the OSPFv3 Inter-Area-Router-LSA. The presence of the base LSA is not mandatory for the usage of the extended LSA with the OSPF FAAM Sub-TLV. This means that the order in which these LSAs are received is not significant.

11. Advertisement of Node Participation in a Flex-Algorithm

When a router is configured to support a particular Flex-Algorithm, we say it is participating in that Flex-Algorithm.

Paths computed for a specific Flex-Algorithm MAY be used by various applications, each potentially using its own specific data plane for forwarding traffic over such paths. To guarantee the presence of the application specific forwarding state associated with a particular Flex-Algorithm, a router MUST advertise its participation for a particular Flex-Algorithm for each application specifically.

11.1. Advertisement of Node Participation for Segment Routing

[RFC8667], [RFC8665], and [RFC8666] (IGP Segment Routing extensions) describe how the SR-Algorithm is used to compute the IGP best path.

Routers advertise the support for the SR-Algorithm as a node capability as described in the above mentioned IGP Segment Routing extensions. To advertise participation for a particular Flex-Algorithm for Segment Routing, including both SR MPLS and SRv6, the Flex-Algorithm value MUST be advertised in the SR-Algorithm TLV (OSPF) or sub-TLV (ISIS).

Segment Routing Flex-Algorithm participation advertisement is topology independent. When a router advertises participation in an SR-Algorithm, the participation applies to all topologies in which the advertising node participates.

11.2. Advertisement of Node Participation for Other Applications

This section describes considerations related to how other applications can advertise their participation in a specific Flex-Algorithm.

Application-specific Flex-Algorithm participation advertisements MAY be topology specific or MAY be topology independent, depending on the application itself.

Application-specific advertisement for Flex-Algorithm participation MUST be defined for each application and is outside of the scope of this document.

12. Advertisement of Link Attributes for Flex-Algorithm

Various link attributes may be used during the Flex-Algorithm path calculation. For example, include or exclude rules based on link affinities can be part of the Flex-Algorithm definition as defined in Section 6 and Section 7.

Link attribute advertisements that are to be used during Flex-Algorithm calculation MUST use the Application-Specific Link Attribute (ASLA) advertisements defined in [RFC8919] or [RFC8920], unless, in the case of IS-IS, the L-Flag is set in the ASLA advertisement. If the L-Flag is set, as defined in [RFC8919] Section 4.2 subject to the constraints discussed in Section 6 of the [[RFC8919], then legacy advertisements are to be used instead.

The mandatory use of ASLA advertisements applies to link attributes specifically mentioned in this document (Min Unidirectional Link Delay, TE Default Metric, Administrative Group, Extended Administrative Group and Shared Risk Link Group) and any other link attributes that may be used in support of Flex-Algorithm in the future.

A new Application Identifier Bit is defined to indicate that the ASLA advertisement is associated with the Flex-Algorithm application. This bit is set in the Standard Application Bit Mask (SABM) defined in [RFC8919] or [RFC8920]:

Bit-3: Flexible Algorithm (X-bit)

ASLA Admin Group Advertisements to be used by the Flexible Algorithm Application MAY use either the Administrative Group or Extended Administrative Group encodings. If the Administrative Group encoding is used, then the first 32 bits of the corresponding FAD sub-TLVs are mapped to the link attribute advertisements as specified in RFC 7308.

13. Calculation of Flexible Algorithm Paths

A router MUST be configured to participate in a given Flex-Algorithm K and MUST select the FAD based on the rules defined in Section 5.3 before it can compute any path for that Flex-Algorithm.

As described in Section 11, participation for any particular Flex-Algorithm MUST be advertised on a per-application basis. Calculation of the paths for any particular Flex-Algorithm MUST be application specific.

The way applications handle nodes that do not participate in Flexible-Algorithm is application specific. If the application only wants to consider participating nodes during the Flex-Algorithm calculation, then when computing paths for a given Flex-Algorithm, all nodes that do not advertise participation for that Flex-Algorithm in their application-specific advertisements MUST be pruned from the topology. Segment Routing, including both SR MPLS and SRv6, are applications that MUST use such pruning when computing Flex-Algorithm paths.

When computing the path for a given Flex-Algorithm, the metric-type that is part of the Flex-Algorithm definition (Section 5) MUST be used.

When computing the path for a given Flex-Algorithm, the calculation-type that is part of the Flex-Algorithm definition (Section 5) MUST be used.

Various link include or exclude rules can be part of the Flex-Algorithm definition. To refer to a particular bit within an AG or EAG we use the term 'color'.

Rules, in the order as specified below, MUST be used to prune links from the topology during the Flex-Algorithm computation.

For all links in the topology:

1. Check if any exclude AG rule is part of the Flex-Algorithm definition. If such exclude rule exists, check if any color that is part of the exclude rule is also set on the link. If such a color is set, the link MUST be pruned from the computation.

2. Check if any exclude SRLG rule is part of the Flex-Algorithm definition. If such exclude rule exists, check if the link is part of any SRLG that is also part of the SRLG exclude rule. If the link is part of such SRLG, the link MUST be pruned from the computation.

3. Check if any include-any AG rule is part of the Flex-Algorithm definition. If such include-any rule exists, check if any color that is part of the include-any rule is also set on the link. If no such color is set, the link MUST be pruned from the computation.

4. Check if any include-all AG rule is part of the Flex-Algorithm definition. If such include-all rule exists, check if all colors that are part of the include-all rule are also set on the link. If all such colors are not set on the link, the link MUST be pruned from the computation.

5. If the Flex-Algorithm definition uses other than IGP metric (Section 5), and such metric is not advertised for the particular link in a topology for which the computation is done, such link

MUST be pruned from the computation. A metric of value 0 MUST NOT be assumed in such case.

13.1. Multi-area and Multi-domain Considerations

Any IGP Shortest Path Tree calculation is limited to a single area. This applies to Flex-Algorithm calculations as well. Given that the computing router does not have visibility of the topology of the next areas or domain, the Flex-Algorithm specific path to an inter-area or inter-domain prefix will be computed for the local area only. The egress L1/L2 router (ABR in OSPF), or ASBR for inter-domain case, will be selected based on the best path for the given Flex-Algorithm in the local area and such egress ABR or ASBR router will be responsible to compute the best Flex-Algorithm specific path over the next area or domain. This may produce an end-to-end path, which is sub-optimal based on Flex-Algorithm constraints. In cases where the ABR or ASBR has no reachability to a prefix for a given Flex-Algorithm in the next area or domain, the traffic may be dropped by the ABR/ASBR.

To allow the optimal end-to-end path for an inter-area or inter-domain prefix for any Flex-Algorithm to be computed, the FAPM has been defined in Section 8 and Section 9. For external route calculation for prefixes originated by ASBRs in remote areas in OSPF, the FAAM has been defined in Section 10.2 for the ABR to indicate its ASBR reachability along with the metric for the specific Flex-Algorithm.

If the FAD selected based on the rules defined in Section 5.3 includes the M-flag, an ABR or ASBR MUST include the FAPM (Section 8, Section 9) when advertising the prefix, that is reachable in a given Flex-Algorithm, between areas or domains. Such metric will be equal to the metric to reach the prefix for that Flex-Algorithm in its source area or domain. This is similar in nature to how the metric is set when prefixes are advertised between areas or domains for the default algorithm. When a prefix is unreachable in its source area or domain in a specific Flex-Algorithm, then an ABR or ASBR MUST NOT include the FAPM for that Flex-Algorithm when advertising the prefix between areas or domains.

If the FAD selected based on the rules defined in Section 5.3 includes the M-flag, the FAPM MUST be used during the calculation of prefix reachability for the inter-area and external prefixes. If the FAPM for the Flex-Algorithm is not advertised with the inter-area or external prefix reachability advertisement, the prefix MUST be considered as unreachable for that Flex-Algorithm. Similarly in the case of OSPF, for ASBRs in remote areas, if the FAAM is not advertised by the local ABR(s), the ASBR MUST be considered as unreachable for that Flex-Algorithm and the external prefix advertisements from such an ASBR are not considered for that Flex-Algorithm.

Flex-Algorithm prefix metrics and the OSPF Flex-Algorithm ASBR metrics MUST NOT be used during the Flex-Algorithm computation unless the FAD selected based on the rules defined in Section 5.3 includes the M-Flag, as described in (Section 6.4 or Section 7.4).

In the case of OSPF, when calculating external routes in a Flex-Algorithm (with FAD selected includes the M-Flag) where the advertising ASBR is in a remote area, the metric will be the sum of the following:

- o the FAPM for that Flex-Algorithm advertised with the external route by the ASBR

- o the metric to reach the ASBR for that Flex-Algorithm from the local ABR i.e., the FAAM for that Flex-Algorithm advertised by the ABR in the local area for that ASBR

- o the Flex-Algorithm specific metric to reach the local ABR

This is similar in nature to how the metric is calculated for routes learned from remote ASBRs in the default algorithm using the OSPFv2 Type 4 ASBR Summary LSA and the OSPFv3 Inter-Area-Router LSA.

If the FAD selected based on the rules defined in Section 5.3 does not includes the M-flag, then the IGP metrics associated with the prefix reachability advertisements used by the base ISIS and OSPF protocol MUST be used for the Flex-Algorithm route computation. Similarly, in the case of external route calculations in OSPF, the ASBR reachability is determined based on the base OSPFv2 Type 4 Summary LSA and the OSFPv3 Inter-Area-Router LSA.

It is NOT RECOMMENDED to use the Flex-Algorithm for inter-area or inter-domain prefix reachability without the M-flag set. The reason is that without the explicit Flex-Algorithm Prefix Metric advertisement (and the Flex-Algorithm ASBR metric advertisement in the case of OSPF external route calculation), it is not possible to conclude whether the ABR or ASBR has reachability to the inter-area or inter-domain prefix for a given Flex-Algorithm in the next area or domain. Sending the Flex-Algoritm traffic for such prefix towards the ABR or ASBR may result in traffic looping or black-holing.

During the route computation, it is possible for the Flex-Algorithm specific metric to exceed the maximum value that can be stored in an unsigned 32-bit variable. In such scenarios, the value MUST be considered to be of value 4,294,967,295 during the computation and advertised as such.

The FAPM MUST NOT be advertised with ISIS L1 or L2 intra-area, OSPFv2 intra-area, or OSPFv3 intra-area routes. If the FAPM is advertised for these route-types, it MUST be ignored during the prefix reachability calculation.

The M-flag in FAD is not applicable to prefixes advertised as SRv6 locators. The ISIS SRv6 Locator TLV [I-D.ietf-lsr-isis-srv6-extensions] includes the Algorithm and Metric fields. When the SRv6 Locator is advertised between areas or domains, the metric field in the Locator TLV of ISIS MUST be used irrespective of the M-flag in the FAD advertisement.

OSPF external and NSSA external prefix advertisements MAY include a non-zero forwarding address in the prefix advertisements in the base protocol. In such a scenario, the Flex-Algorithm specific reachability of the external prefix is determined by Flex-Algorithm specific reachability of the forwarding address.

In OSPF, the procedures for translation of NSSA external prefix advertisements into external prefix advertisements performed by an NSSA ABR [RFC3101] remain unchanged for Flex-Algorithm. An NSSA translator MUST include the OSPF FAPM Sub-TLVs for all Flex-Algorithms that are in the original NSSA external prefix advertisement from the NSSA ASBR in the translated external prefix advertisement generated by it regardless of its participation in those Flex-Algorithms or its having reachability to the NSSA ASBR in those Flex-Algorithms.

14. Flex-Algorithm and Forwarding Plane

This section describes how Flex-Algorithm paths are used in forwarding.

14.1. Segment Routing MPLS Forwarding for Flex-Algorithm

This section describes how Flex-Algorithm paths are used with SR MPLS forwarding.

Prefix SID advertisements include an SR-Algorithm value and, as such, are associated with the specified SR-Algorithm. Prefix-SIDs are also associated with a specific topology which is inherited from the associated prefix reachability advertisement. When the algorithm value advertised is a Flex-Algorithm value, the Prefix SID is associated with paths calculated using that Flex-Algorithm in the associated topology.

A Flex-Algorithm path MUST be installed in the MPLS forwarding plane using the MPLS label that corresponds to the Prefix-SID that was advertised for that Flex-algorithm. If the Prefix SID for a given Flex-algorithm is not known, the Flex-Algorithm specific path cannot be installed in the MPLS forwarding plane.

Traffic that is supposed to be routed via Flex-Algorithm specific paths, MUST be dropped when there are no such paths available.

Loop Free Alternate (LFA) paths for a given Flex-Algorithm MUST be computed using the same constraints as the calculation of the primary paths for that Flex-Algorithm. LFA paths MUST only use Prefix-SIDs advertised specifically for the given algorithm. LFA paths MUST NOT use an Adjacency-SID that belongs to a link that has been pruned from the Flex-Algorithm computation.

If LFA protection is being used to protect a given Flex-Algorithm paths, all routers in the area participating in the given Flex-Algorithm SHOULD advertise at least one Flex-Algorithm specific Node-SID. These Node-SIDs are used to steer traffic over the LFA computed backup path.

14.2. SRv6 Forwarding for Flex-Algorithm

This section describes how Flex-Algorithm paths are used with SRv6 forwarding.

In SRv6 a node is provisioned with topology/algorithm specific locators for each of the topology/algorithm pairs supported by that node. Each locator is an aggregate prefix for all SIDs provisioned on that node which have the matching topology/algorithm.

The SRv6 locator advertisement in ISIS [I-D.ietf-lsr-isis-srv6-extensions] includes the MTID value that associates the locator with a specific topology. SRv6 locator advertisements also includes an Algorithm value that explicitly associates the locator with a specific algorithm. When the algorithm value advertised with a locator represents a Flex-Algorithm, the paths to the locator prefix MUST be calculated using the specified Flex-Algorithm in the associated topology.

Forwarding entries for the locator prefixes advertised in ISIS MUST be installed in the forwarding plane of the receiving SRv6 capable routers when the associated topology/algorithm is participating in them. Forwarding entries for locators associated with Flex-Algorithms in which the node is not participating MUST NOT be installed in the forwarding plane.

When the locator is associated with a Flex-Algorithm, LFA paths to the locator prefix MUST be calculated using such Flex-Algorithm in the associated topology, to guarantee that they follow the same constraints as the calculation of the primary paths. LFA paths MUST only use SRv6 SIDs advertised specifically for the given Flex-Algorithm.

If LFA protection is being used to protect locators associated with a given Flex-Algorithm, all routers in the area participating in the given Flex-Algorithm SHOULD advertise at least one Flex-Algorithm specific locator and END SID per node and one END.X SID for every link that has not been pruned from such Flex-Algorithm computation. These locators and SIDs are used to steer traffic over the LFA-computed backup path.

14.3. Other Applications' Forwarding for Flex-Algorithm

Any application that wants to use Flex-Algorithm specific forwarding needs to install some form of Flex-Algorithm specific forwarding entries.

Application-specific forwarding for Flex-Algorithm MUST be defined for each application and is outside of the scope of this document.

15. Operational Considerations

15.1. Inter-area Considerations

The scope of the FA computation is an area, so is the scope of the FAD. In ISIS, the Router Capability TLV in which the FAD Sub-TLV is advertised MUST have the S-bit clear, which prevents it to be flooded outside of the level in which it was originated. Even though in OSPF the FAD Sub-TLV can be flooded in an RI LSA that has AS flooding scope, the FAD selection is performed for each individual area in which it is being used.

There is no requirement for the FAD for a particular Flex-Algorithm to be identical in all areas in the network. For example, traffic for the same Flex-Algorithm may be optimized for minimal delay (e.g., using delay metric) in one area or level, while being optimized for available bandwidth (e.g., using IGP metric) in another area or level.

As described in Section 5.1, ISIS allows the re-generation of the winning FAD from level 2, without any modification to it, into a level 1 area. This allows the operator to configure the FAD in one or multiple routers in the level 2, without the need to repeat the same task in each level 1 area, if the intent is to have the same FAD for the particular Flex-Algorithm across all levels. This can similarly be achieved in OSPF by using the AS flooding scope of the RI LSA in which the FAD Sub-TLV for the particular Flex-Algoritm is advertised.

Re-generation of FAD from a level 1 area to the level 2 area is not supported in ISIS, so if the intent is to regenerate the FAD between ISIS levels, the FAD MUST be defined on router(s) that are in level 2. In OSPF, the FAD definition can be done in any area and be propagated to all routers in the OSPF routing domain by using the AS flooding scope of the RI LSA.

15.2. Usage of SRLG Exclude Rule with Flex-Algorithm

There are two different ways in which SRLG information can be used with Flex-Algorithm:

In a context of a single Flex-Algorithm, it can be used for computation of backup paths, as described in [I-D.ietf-rtgwg-segment-routing-ti-lfa]. This usage does not require association of any specific SRLG constraint with the given Flex-Algorithm definition.

In the context of multiple Flex-Algorithms, it can be used for creating disjoint sets of paths by pruning the links belonging to a specific SRLG from the topology on which a specific Flex-Algorithm computes its paths. This usage:

Facilitates the usage of already deployed SRLG configurations for setup of disjoint paths between two or more Flex-Algorithms.

Requires explicit association of a given Flex-Algorithm with a specific set of SRLG constraints as defined in Section 6.5 and Section 7.5.

The two usages mentioned above are orthogonal.

15.3. Max-metric consideration

Both ISIS and OSPF have a mechanism to set the IGP metric on a link to a value that would make the link either non-reachable or to serve as the link of last resort. Similar functionality would be needed for the Min Unidirectional Link Delay and TE metric, as these can be used to compute Flex-Algorithm paths.

The link can be made un-reachable for all Flex-Algorithms that use Min Unidirectional Link Delay as metric, as described in Section 5.1, by removing the Flex-Algorithm ASLA Min Unidirectional Link Delay advertisement for the link. The link can be made the link of last resort by setting the delay value in the Flex-Algorithm ASLA delay advertisement for the link to the value of 16,777,215 (2^24 - 1).

The link can be made un-reachable for all Flex-Algorithms that use TE metric, as described in Section 5.1, by removing the Flex-Algorithm ASLA TE metric advertisement for the link. The link can be made the link of last resort by setting the TE metric value in the Flex-Algorithm ASLA delay advertisement for the link to the value of (2^24 - 1) in ISIS and (2^32 - 1) in OSPF.

16. Backward Compatibility

This extension brings no new backward compatibility issues. ISIS, OSPFv2 and OSPFv3 all have well defined handling of unrecognized TLVs and sub-TLVs that allows the introduction of the new extensions, similar to those defined here, without introducing any interoperability issues.

17. Security Considerations

This draft adds two new ways to disrupt IGP networks:

> An attacker can hijack a particular Flex-Algorithm by advertising a FAD with a priority of 255 (or any priority higher than that of the legitimate nodes).
>
> An attacker could make it look like a router supports a particular Flex-Algorithm when it actually doesn't, or vice versa.

Both of these attacks can be addressed by the existing security extensions as described in [RFC5304] and [RFC5310] for ISIS, in [RFC2328] and [RFC7474] for OSPFv2, and in [RFC5340] and [RFC4552] for OSPFv3.

18. IANA Considerations

18.1. IGP IANA Considerations

18.1.1. IGP Algorithm Types Registry

This document makes the following registrations in the "IGP Algorithm Types" registry:

> Type: 128-255.
>
> Description: Flexible Algorithms.

Reference: This document (Section 4).

18.1.2. IGP Metric-Type Registry

IANA is requested to set up a registry called "IGP Metric-Type Registry" under an "Interior Gateway Protocol (IGP) Parameters" IANA registries. The registration policy for this registry is "Standards Action" ([RFC8126] and [RFC7120]).

Values in this registry come from the range 0-255.

This document registers following values in the "IGP Metric-Type Registry":

Type: 0

Description: IGP metric

Reference: This document (Section 5.1)

Type: 1

Description: Min Unidirectional Link Delay as defined in [RFC8570], section 4.2, and [RFC7471], section 4.2.

Reference: This document (Section 5.1)

Type: 2

Description: Traffic Engineering Default Metric as defined in [RFC5305], section 3.7, and Traffic engineering metric as defined in [RFC3630], section 2.5.5

Reference: This document (Section 5.1)

18.2. Flexible Algorithm Definition Flags Registry

IANA is requested to set up a registry called "ISIS Flexible Algorithm Definition Flags Registry" under an "Interior Gateway Protocol (IGP) Parameters" IANA registries. The registration policy for this registry is "Standards Action" ([RFC8126] and [RFC7120]).

This document defines the following single bit in Flexible Algorithm Definition Flags registry:

Bit #	Name
0	Prefix Metric Flag (M-flag)

Reference: This document (Section 6.4, Section 7.4).

18.3. ISIS IANA Considerations

18.3.1. Sub TLVs for Type 242

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

Type: 26.

Description: Flexible Algorithm Definition.

Reference: This document (Section 5.1).

18.3.2. Sub TLVs for for TLVs 135, 235, 236, and 237

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

Type: 6

Description: Flexible Algorithm Prefix Metric.

Reference: This document (Section 8).

18.3.3. Sub-Sub-TLVs for Flexible Algorithm Definition Sub-TLV

This document creates the following Sub-Sub-TLV Registry:

Registry: Sub-Sub-TLVs for Flexible Algorithm Definition Sub-TLV

Registration Procedure: Expert review

Reference: This document (Section 5.1)

This document defines the following Sub-Sub-TLVs in the "Sub-Sub-TLVs for Flexible Algorithm Definition Sub-TLV" registry:

Type: 1

Description: Flexible Algorithm Exclude Admin Group

Reference: This document (Section 6.1).

Type: 2

Description: Flexible Algorithm Include-Any Admin Group

Reference: This document (Section 6.2).

Type: 3

Description: Flexible Algorithm Include-All Admin Group

Reference: This document (Section 6.3).

Type: 4

Description: Flexible Algorithm Definition Flags

Reference: This document (Section 6.4).

Type: 5

Description: Flexible Algorithm Exclude SRLG

Reference: This document (Section 6.5).

18.4. OSPF IANA Considerations

18.4.1. OSPF Router Information (RI) TLVs Registry

This specification updates the OSPF Router Information (RI) TLVs Registry.

Type: 16

Description: Flexible Algorithm Definition TLV.

Reference: This document (Section 5.2).

18.4.2. OSPFv2 Extended Prefix TLV Sub-TLVs

This document makes the following registrations in the "OSPFv2 Extended Prefix TLV Sub-TLVs" registry.

Type: 3

Description: Flexible Algorithm Prefix Metric.

Reference: This document (Section 9).

18.4.3. OSPFv3 Extended-LSA Sub-TLVs

This document makes the following registrations in the "OSPFv3 Extended-LSA Sub-TLVs" registry.

Type: 26

Description: Flexible Algorithm Prefix Metric.

Reference: This document (Section 9).

Type: TBD (suggested value 30)

Description: OSPF Flexible Algorithm ASBR Metric Sub-TLV

Reference: This document (Section 10.2).

18.4.4. OSPF Flex-Algorithm Prefix Metric Bits

This specification requests creation of "OSPF Flex-Algorithm Prefix Metric Bits" registry under the OSPF Parameters Registry with the following initial values.

Bit Number: 0

Description: E bit - External Type

Reference: this document.

The bits 1-7 are unassigned and the registration procedure to be followed for this registry is IETF Review.

18.4.5. OSPF Opaque LSA Option Types

This document makes the following registrations in the "OSPF Opaque LSA Option Types" registry.

Value: TBD (suggested value 11)

Description: OSPFv2 Extended Inter-Area ASBR LSA

Reference: This document (Section 10.1).

18.4.6. OSPFv2 Externded Inter-Area ASBR TLVs

This specification requests creation of "OSPFv2 Extended Inter-Area ASBR TLVs" registry under the OSPFv2 Parameters Registry with the following initial values.

Value: 1

Description : Extended Inter-Area ASBR TLV

Reference: this document

The values 2 to 32767 are unassigned, values 32768 to 33023 are reserved for experimental use while the values 0 and 33024 to 65535 are reserved. The registration procedure to be followed for this registry is IETF Review or IESG Approval.

18.4.7. OSPFv2 Inter-Area ASBR Sub-TLVs

This specification requests creation of "OSPFv2 Extended Inter-Area ASBR Sub-TLVs" registry under the OSPFv2 Parameters Registry with the following initial values.

Value: 1

Description : OSPF Flexible Algorithm ASBR Metric Sub-TLV

Reference: this document

The values 2 to 32767 are unassigned, values 32768 to 33023 are reserved for experimental use while the values 0 and 33024 to 65535 are reserved. The registration procedure to be followed for this registry is IETF Review or IESG Approval.

18.4.8. OSPF Flexible Algorithm Definition TLV Sub-TLV Registry

This document creates the following registry:

Registry: OSPF Flexible Algorithm Definition TLV sub-TLV

Registration Procedure: Expert review

Reference: This document (Section 5.2)

The "OSPF Flexible Algorithm Definition TLV sub-TLV" registry will define sub-TLVs at any level of nesting for the Flexible Algorithm TLV and should be added to the "Open Shortest Path First (OSPF)

Parameters" registries group. New values can be allocated via IETF Review or IESG Approval.

This document registers following Sub-TLVs in the "TLVs for Flexible Algorithm Definition TLV" registry:

Type: 1

Description: Flexible Algorithm Exclude Admin Group

Reference: This document (Section 7.1).

Type: 2

Description: Flexible Algorithm Include-Any Admin Group

Reference: This document (Section 7.2).

Type: 3

Description: Flexible Algorithm Include-All Admin Group

Reference: This document (Section 7.3).

Type: 4

Description: Flexible Algorithm Definition Flags

Reference: This document (Section 7.4).

Type: 5

Description: Flexible Algorithm Exclude SRLG

Reference: This document (Section 7.5).

Types in the range 32768-33023 are for experimental use; these will not be registered with IANA, and MUST NOT be mentioned by RFCs.

Types in the range 33024-65535 are not to be assigned at this time. Before any assignments can be made in the 33024-65535 range, there MUST be an IETF specification that specifies IANA Considerations that covers the range being assigned.

18.4.9. Link Attribute Applications Registry

This document registers following bit in the Link Attribute Applications Registry:

Bit-3

Description: Flexible Algorithm (X-bit)

Reference: This document (Section 12).

19. Acknowledgements

This draft, among other things, is also addressing the problem that the [I-D.gulkohegde-routing-planes-using-sr] was trying to solve. All authors of that draft agreed to join this draft.

Thanks to Eric Rosen, Tony Przygienda, William Britto A J, Gunter Van De Velde, Dirk Goethals, Manju Sivaji and, Baalajee S for their detailed review and excellent comments.

Thanks to Cengiz Halit for his review and feedback during initial phase of the solution definition.

Thanks to Kenji Kumaki for his comments.

Thanks to Acee Lindem for editorial comments.

20. References

20.1. Normative References

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extension to Support Segment Routing over IPv6 Dataplane", draft-ietf-lsr-isis-srv6-extensions-14 (work in progress), April 2021.

[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>.

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <https://www.rfc-editor.org/info/rfc4203>.

[RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, DOI 10.17487/RFC5250, July 2008, <https://www.rfc-editor.org/info/rfc5250>.

[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, <https://www.rfc-editor.org/info/rfc5307>.

[RFC7308] Osborne, E., "Extended Administrative Groups in MPLS Traffic Engineering (MPLS-TE)", RFC 7308, DOI 10.17487/RFC7308, July 2014, <https://www.rfc-editor.org/info/rfc7308>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

[RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <https://www.rfc-editor.org/info/rfc7770>.

[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>.

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8666] Psenak, P., Ed. and S. Previdi, Ed., "OSPFv3 Extensions for Segment Routing", RFC 8666, DOI 10.17487/RFC8666, December 2019, <https://www.rfc-editor.org/info/rfc8666>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

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

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

20.2. Informative References

[I-D.gulkohegde-routing-planes-using-sr]
Hegde, S. and a. arkadiy.gulko@thomsonreuters.com, "Separating Routing Planes using Segment Routing", draft-gulkohegde-routing-planes-using-sr-00 (work in progress), March 2017.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]
Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-06 (work in progress), February 2021.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC3101] Murphy, P., "The OSPF Not-So-Stubby Area (NSSA) Option", RFC 3101, DOI 10.17487/RFC3101, January 2003, <https://www.rfc-editor.org/info/rfc3101>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, DOI 10.17487/RFC3906, October 2004, <https://www.rfc-editor.org/info/rfc3906>.

[RFC4552] Gupta, M. and N. Melam, "Authentication/Confidentiality for OSPFv3", RFC 4552, DOI 10.17487/RFC4552, June 2006, <https://www.rfc-editor.org/info/rfc4552>.

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

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

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

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC7120] Cotton, M., "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 7120, DOI 10.17487/RFC7120, January 2014, <https://www.rfc-editor.org/info/rfc7120>.

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

[RFC7474] Bhatia, M., Hartman, S., Zhang, D., and A. Lindem, Ed., "Security Extension for OSPFv2 When Using Manual Key Management", RFC 7474, DOI 10.17487/RFC7474, April 2015, <https://www.rfc-editor.org/info/rfc7474>.

[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>.

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

Authors' Addresses

Peter Psenak (editor)
Cisco Systems
Apollo Business Center
Mlynske nivy 43
Bratislava, 82109
Slovakia

Email: ppsenak@cisco.com

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

Email: shraddha@juniper.net

Clarence Filsfils
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Ketan Talaulikar
Cisco Systems, Inc.
S.No. 154/6, Phase I, Hinjawadi
PUNE, MAHARASHTRA 411 057
India

Email: ketant@cisco.com

Arkadiy Gulko
Edward Jones

Email: arkadiy.gulko@edwardjones.com

LSR Shaofu. Peng
Internet-Draft Ran. Chen
Intended status: Standards Track ZTE Corporation
Expires: June 10, 2021 Ketan. Talaulikar
Cisco Systems
December 7, 2020

Algorithm Related IGP-Adjacency SID Advertisement
draft-peng-lsr-algorithm-related-adjacency-sid-02

Abstract

Segment Routing architecture supports the use of multiple routing algorithms, i.e, different constraint-based shortest-path calculations can be supported. There are two standard algorithms: SPF and Strict-SPF, defined in Segment Routing architecture. There are also other user defined algorithms according to Flex-algo applicaiton. However, an algorithm identifier is often included as part of a Prefix-SID advertisement, that maybe not satisfy some scenarios where multiple algorithm share the same link resource. This document complement that the algorithm identifier can be also included as part of a Adjacency-SID advertisement

Table of Contents

1. Introduction

Segment Routing architecture [RFC8402] supports the use of multiple routing algorithms, i.e, different constraint-based shortest-path calculations can be supported. There are two standard algorithms, i.e, SPF and Strict-SPF, that defined in Segment Routing architecture. For SPF, the packet is forwarded along the well known ECMP-aware Shortest Path First (SPF) algorithm employed by the IGPs. However, it is explicitly allowed for a midpoint to implement another forwarding based on local policy. For Strict Shortest Path First (Strict-SPF), it mandates that the packet be forwarded according to the ECMP-aware SPF algorithm and instructs any router in the path to ignore any possible local policy overriding the SPF decision.

There are also other user defined algorithms according to IGP Flex Algorithm [I-D.ietf-lsr-flex-algo]. IGP Flex Algorithm proposes a solution that allows IGPs themselves to compute constraint based paths over the network, and it also specifies a way of using Segment Routing (SR) Prefix-SIDs and SRv6 locators to steer packets along the constraint-based paths. It specifies a set of extensions to ISIS, OSPFv2 and OSPFv3 that enable a router to send TLVs that identify (a) calculation-type, (b) specify a metric-type, and (c)describe a set of constraints on the topology, that are to be used to compute the best paths along the constrained topology. A given combination of calculation-type, metric-type, and constraints is known as an FAD (Flexible Algorithm Definition).

However, an algorithm identifier is often included as part of a Prefix-SID advertisement, that maybe not satisfy some scenarios where multiple algorithm share the same link resource. For example, an SR-TE policy may be instantiated within specific Flex-algo plane, i.e., the SID list requires to include algorithm related SIDs. An algorithm-unware Adjacency-SID included in the SID list can just steer the packet towards the link, but can not apply different QoS policy for different algorithm. Another example is that the TI-LFA backup path computed in Flex-algo plane may also contain an algorithm-unware Adjacency-SID, which maybe also used in other SR-TE instance that carries other service.

This document complement that the algorithm identifier can be also included as part of an Adjacency-SID advertisement for SR-MPLS.

2. Requirements Language

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

3. Adjacency Segment Identifier per Algorithm

3.1. ISIS Adjacency Segment Identifier per Algorithm

[RFC8667] describes the IS-IS extensions that need to be introduced for Segment Routing operating on an MPLS data plane. It defined Adjacency Segment Identifier (Adj-SID) sub-TLV advertised with TLV-22/222/23/223/141, and Adjacency Segment Identifier (LAN-Adj-SID) Sub-TLV advetised with TLV-22/222/23/223. Accordingly, this document defines two new optional Sub-TLVs, "ISIS Adjacency Segment Identifier (Adj-SID) per Algorithm Sub-TLV" and "ISIS Adjacency Segment Identifier (LAN-Adj-SID) per Algorithm Sub-TLV".

3.1.1. ISIS Adjacency Segment Identifier (Adj-SID) per Algorithm Sub-TLV

ISIS Adjacency Segment Identifier (Adj-SID) per Algorithm Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |     Length    |     Flags     |     Weight    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Algorithm   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          SID/Label/Index (variable)           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: ISIS Adjacency Segment Identifier (Adj-SID) per Algorithm Format

where:

Type: TBD1.

Length: 6 or 7 depending on size of the SID.

Flags: Refer to Adjacency Segment Identifier (Adj-SID) sub-TLV.

Weight: Refer to Adjacency Segment Identifier (Adj-SID) sub-TLV.

Algorithm: The Algorithm field contains the identifier of the algorithm the router uses to apply algorithm specific QoS policy configured on the adjacency.

SID/Label/Index: Refer to Adjacency Segment Identifier (Adj-SID) sub-TLV.

For a P2P link, an SR-capable router MAY allocate different Adj-SID for different algorithm, if this link will join different algorithm related plane.

3.1.2. ISIS Adjacency Segment Identifier (LAN-Adj-SID) per Algorithm Sub-TLV

ISIS Adjacency Segment Identifier (LAN-Adj-SID) per Algorithm Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |     Length    |     Flags     |    Weight     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Algorithm   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Neighbor System-ID (ID length octets)        |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   SID/Label/Index (variable)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: ISIS Adjacency Segment Identifier (LAN-Adj-SID) per Algorithm Format

where:

Type: TBD2.

Length: Variable.

Flags: Refer to Adjacency Segment Identifier (LAN-Adj-SID) Sub-TLV.

Weight: Refer to Adjacency Segment Identifier (LAN-Adj-SID) Sub-TLV.

Algorithm: The Algorithm field contains the identifier of the algorithm the router uses to apply algorithm specific QoS policy configured on the adjacency.

SID/Label/Index: Refer to Adjacency Segment Identifier (LAN-Adj-SID) Sub-TLV.

For a broadcast link, an SR-capable router MAY allocate different Adj-SID for different algorithm, if this link will join different algorithm related plane.

3.2. OSPF Adjacency Segment Identifier per Algorithm

[RFC8665] describes the OSPF extensions that need to be introduced for Segment Routing operating on an MPLS data plane. It defined Adj-SID Sub-TLV and LAN Adj-SID Sub-TLV advertised with Extended Link TLV defined in [RFC7684]. This document extends these two Sub-TLVs to carry the specific algorithm.

3.2.1. OSPF Adj-SID Sub-TLV

The existing Adj-SID Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     |   Algorithm   |  MT-ID        |  Weight       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   SID/Label/Index (variable)                  |
+---------------------------------------------------------------+
```

Figure 3: OSPF Adj-SID Format

where:

Algorithm: The new Algorithm field contains the identifier of the algorithm the router uses to apply algorithm specific QoS policy configured on the adjacency.

For a P2P link, an SR-capable router MAY allocate different Adj-SID for different algorithm, if this link will join different algorithm related plane.

3.2.2. OSPF LAN Adj-SID Sub-TLV

The existing LAN Adj-SID Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     |   Algorithm   |     MT-ID     |    Weight     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Neighbor ID                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    SID/Label/Index (variable)                 |
+---------------------------------------------------------------+
```

Figure 4: OSPF LAN Adj-SID Format

where:

Algorithm: The new Algorithm field contains the identifier of the algorithm the router uses to apply algorithm specific QoS policy configured on the adjacency.

For a broadcast link, an SR-capable router MAY allocate different Adj-SID for different algorithm, if this link will join different algorithm related plane.

3.3. OSPFv3 Adjacency Segment Identifier per Algorithm

[RFC8666] describes the OSPFv3 extensions that need to be introduced for Segment Routing operating on an MPLS data plane. It defined Adj-SID Sub-TLV and LAN Adj-SID Sub-TLV advertised with Router-Link TLV as defined in [RFC8362]. This document extends these two Sub-TLVs to carry the specific algorithm.

3.3.1. OSPFv3 Adj-SID Sub-TLV

The existing Adj-SID Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type               |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Flags         |     Weight      |  Algorithm     |   Reserved     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   SID/Label/Index (variable)                    |
+-----------------------------------------------------------------+
```

Figure 5: OSPFv3 Adj-SID Format

where:

Algorithm: The new Algorithm field contains the identifier of the algorithm the router uses to apply algorithm specific QoS policy configured on the adjacency.

For a P2P link, an SR-capable router MAY allocate different Adj-SID for different algorithm, if this link will join different algorithm related plane.

3.3.2. OSPFv3 LAN Adj-SID Sub-TLV

The existing LAN Adj-SID Sub-TLV has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |              Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     |     Weight    |   Algorithm   |    Reserved   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Neighbor ID                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   SID/Label/Index (variable)                  |
+---------------------------------------------------------------+
```

Figure 6: OSPFv3 LAN Adj-SID Format

where:

Algorithm: The new Algorithm field contains the identifier of the algorithm the router uses to apply algorithm specific QoS policy configured on the adjacency.

For a broadcast link, an SR-capable router MAY allocate different Adj-SID for different algorithm, if this link will join different algorithm related plane.

4. Operations

The method introduced in this document enables the traffic of different flex-algo plane to be distinguished on the same link, so that these traffic can be applied with different QoS policy per algorithm.

The endpoint of a link shared by multiple flex-algo plane can reserve different queue resources for different algorithms locally, and perform priority based queue scheduling and traffic shaping. This algorithm related reserved information can be advertised to other nodes in the network through some mechanism, therefore it has an impact on the constraint based path calculation of the flex-algo plane. How to allocate algorithm related resouce and advertise it in the network is out the scope of this document.

Depending on the implementation, operators can configure multiple Adacency-SIDs each for different algorithm on the same link. One of the difficulties is that during this configuration phase it is not straightforward for a link to be included in an FA plane, as this can only be determined after all nodes in the network have negotiated the FAD. A simple way is that as long as an IGP instance enable an FA for a level/area, all links joined to that level/area should allocate Adjacency-SIDs for that algorithm statically. Another way is to allocate and withdraw Adjacency-SID per algorithm dynamically according to the result of FAD negotiation.

The following figure shows an example of Adjacency-SID per algorithm.

```
[S1]--------[D]--------[S2]
 |           |           |
 |           |           |
 |           |           |
[A]---------[B]--------[C]
```

Figure 7: Flex-algo LFA Path with Adjacency-SID per Algorithm

Suppose that node S1, A, B, D and their inter-connected links belongs to FA-id 128 plane, and S2, B, C, D and their inter-connected links belongs to FA-id 129 plane. The IGP metric of link B-D is 100, and all other links have IGP metric 1. In FA-id 128 plane, from S1 to destination D, the primary path is S1-D, and the TI-LFA backup path is segment list {node(B), adjacency(B-D)}. Similarly, In FA-id 129 plane, from S2 to destination D, the primary path is S2-D, and the TI-LFA backup path is segment list {node(B), adjacency(B-D)}. The above TI-LFA path of FA-id 128 plane can be translated to {node-SID(B)@FA-id128, adjacency-SID(B-D)@FA-id128}, and TI-LFA path of FA-id 129 plane will be translate to {node-SID(B)@FA-id129, adjacency-SID(B-D)@FA-id129}. So that node B can distinguish the flow of FA-id 128 and FA-id 129 based on different adjacency-SID(B-D), and take different treatment (e.g., QoS policy) of them when they are send to the same outgoing link B-D.

5. IANA Considerations

TBD

6. Security Considerations

There are no new security issues introduced by the extensions in this document. Refer to [RFC8665], [RFC8666], [RFC8667] for other security considerations.

7. Acknowledgements

TBD

8. Normative References

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[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>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[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>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

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

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8666] Psenak, P., Ed. and S. Previdi, Ed., "OSPFv3 Extensions for Segment Routing", RFC 8666, DOI 10.17487/RFC8666, December 2019, <https://www.rfc-editor.org/info/rfc8666>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC8668] Ginsberg, L., Ed., Bashandy, A., Filsfils, C., Nanduri, M., and E. Aries, "Advertising Layer 2 Bundle Member Link Attributes in IS-IS", RFC 8668, DOI 10.17487/RFC8668, December 2019, <https://www.rfc-editor.org/info/rfc8668>.

Authors' Addresses

Shaofu Peng
ZTE Corporation
China

Email: peng.shaofu@zte.com.cn

Ran Chen
ZTE Corporation
China

Email: chen.ran@zte.com.cn

Ketan Talaulikar
Cisco Systems
India

Email: ketant@cisco.com

Internet-Draft
Intended status: Standards Track
Expires: June 10, 2021

Y. Zhu
China telecom
S. Peng
R. Chen
ZTE Corporation
G. Mirsky
ZTE Corp.
December 7, 2020

IGP Flexible Algorithm with L2bundles
draft-peng-lsr-flex-algo-l2bundles-05

Abstract

IGP Flex Algorithm proposes a solution that allows IGPs themselves to compute constraint based paths over the network, and it also specifies a way of using Segment Routing (SR) Prefix-SIDs and SRv6 locators to steer packets along the constraint-based paths. This document describes how to create Flex-algo plane with L2bundles scenario.

Table of Contents

1. Introduction

IGP Flex Algorithm [I-D.ietf-lsr-flex-algo] proposes a solution that allows IGPs themselves to compute constraint based paths over the network, and it also specifies a way of using Segment Routing [RFC8402] Prefix-SIDs and SRv6 locators to steer packets along the constraint-based paths. It specifies a set of extensions to ISIS, OSPFv2 and OSPFv3 that enable a router to send TLVs that identify (a) calculation-type, (b) specify a metric-type, and (c)describe a set of constraints on the topology, that are to be used to compute the best paths along the constrained topology. A given combination of calculation-type, metric-type, and constraints is known as an FAD (Flexible Algorithm Definition).

[RFC8668] and [I-D.ketant-lsr-ospf-l2bundles] introduces the ability for IS-IS and OSPF respectively to advertise the link attributes of Layer 2 (L2) Bundle Members. Especially, the link attribute "Administrative Group" and "Extended Administrative Group" could be individual to each L2 Bundle Member for purpose of Flex-algo plane construction, where multiple Flex-algo planes share the same Layer 3 parent interface and each Flex-algo plane has dedicated L2 Bundle Member.

This document describes how to create Flex-algo plane with L2bundles scenario.

2. Requirements Language

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

3. Color set on L2 Bundle Member

Traffic Engineering affinity (also termed as Color) is often to be set on the Layer 3 interface and be flooded by IGP-TE. However, when the Layer 3 interface is a Layer 2 interface bundle, operators can config individual color for each L2 Bundle Member. So that IGP link-state database will contain the TE affinity attribute of L2 Bundle Member, as well as Layer 3 parrent interface.

Note that Layer 3 interface can join to IGP instance explicitly, but L2 Bundle Member not.

The TE affinity of the Layer 3 parrent interface can be a combined value of all L2 Bundle Members. For example, if the Layer 3 parrent interface contains three L2 Bundle Members, each with color "RED", "GREEN", "BLUE" respectively, the Layer 3 parrent interface will have color "RED|GREEN|BLUE".

4. Flex-algo plane with L2 link resource

4.1. Best-effort

[I-D.ietf-lsr-flex-algo] defines the color-based link resource selection rules in FAD to construct the expected Flex-algo plane. Each node in the Flex-algo plane will maintain the best path to other destination nodes. In the case of L2bundles scenario, each node need check the outgoing Layer 2 bundle interface, to see which L2 Bundle Member does exactly belong to the Flex-algo plane.

For the node who originate the l2-bundle interface, the forwarding information of the FIB entry with outgoing Layer 2 bundle interface will exactly select the L2 Bundle Member that belongs to the Flex-algo plane to forward packets.

For example, three Flex-algo plane share the same Layer 3 parrent interface including three L2 Bundle Members each with color "RED", "GREEN", "BLUE" respectively, and each Flex-algo plane with link selection rule "Include-Any RED", "Include-Any GREEN", "Include-Any BLUE" respectively, Flex-algo SHOULD NOT simply select the Layer 3 parrent interface for all Flex-algo plane, but need continue to select individual L2 Bundle Member for each specific Flex-algo plane. As a reslut, the FIB entry within Flex-algo RED plane will exactly choose the L2 Bundle Members with color "RED" to forward packets, the FIB entry within Flex-algo GREEN plane will exactly choose the L2 Bundle Members with color "GREEN" to forward packets, and the FIB entry within Flex-algo BLUE plane will exactly choose the L2 Bundle Members with color "BLUE" to forward packets.

The above processing is a local optimization for each node who originate l2-bundle interface.

In addition, for a remote node which received l2-bundle advertisement originated from other nodes, if that l2-bundle is in the flex-algo based path to a destination node, it must confirm which L2 Bundle Member belongs to the flex-algo plane and check that L2 Bundle Member really meets the constraints defined in the related FAD. This processing is necessary when Flex-algo is used to optimize SID stack depth for an SR-TE policy, e.g, the SR-TE policy defines TE affinity to select individual L2 Bundle Member and the SID list may contain Adjacency-SID for a specific L2 Bundle Member as described in [RFC8668] and [I-D.ketant-lsr-ospf-l2bundles]. Thus the flex-algo based path must be consistent with the original path of the optimized SR-TE policy, i.e, within the flex-algo plane when each node determine its next-hop towards a destination, the determination must be based on the above confirmation and check of L2 Bundle Members.

4.2. Traffic Engineering

A segment list contains SIDs advertised specifically for the given algorithm is possible, such as an inter-domain path contains multiple Flex-algo domains, a TI-LFA backup path within the Flex-algo plane, or an optimized TE path avoiding congested link within the Flex-algo plane. When the headend or controller compute these SR-TE paths within the specific flex-algo plane, in addition to the algorithm based Prefix-SID towards the loose node, an Adjacency-SID can also be used to strictly steer the packets along the expected L3 link. However, if the L3 link is a l2-bundle interface, it is necessary to see which L2 Bundle Member exactly belongs to the specific Flex-algo plane and use the Adjacency-SID for that member.

[RFC8668] and [I-D.ketant-lsr-ospf-l2bundles] have defined Adjacency-SID for each L2 Bundle Member, that can be used to isolate flows among multiple Flex-algo planes, when these Flex-algo planes share the same Layer 3 parrent interface. A specific Adjacency-SID for a specific L2 Bundle Member can be contained in the SID list of the SR path within the flex-algo plane and steer the packets to that member.

5. Flex-algo L2bundles Use-cases

In some operator's networks, a large number of bundled links are deployed to improve the bandwidth. However, for a specific l2bundle, each member has different capabilities, such as different delay, bandwidth, AG/EAG, etc. When the path of an SR policy needs to go through an Layer 2 interface bundle, operators want to choose the individual member link to meet business requirements. Different SR policy may choose different member links, according to different set of constraints.

When Flex algorithm is enabled in the above networks, even all flex-algo planes share all Layer 2 interface bundles, i.e, all FA planes have the same structure, an important requirement to Flex-algo is that the constraint based computation of Flex-algo must consider how to select member links to meet service's criterias. In addition, different flex-algo planes can also have different structures, with different set of nodes and links, to meet more strict business requirements.

The extended behavior of flex-algo introduced in this document can meet the above requirement, and exactly it is independent with the structure of flex-algo plane.

5.1. Flex-algo L2bundles Examples

Let's describe the requirement with the following example.

```
S======A=====B======C=====D
\                        /
 \___________E__________/
```

Figure 1: Flex-algo L2bundles Example

An SR policy from headend S to endpoint D is created, with color template {min delay}. Suppose the macthed link is the upper member link of l2bundles interface between S-A, A-B, B-C, C-D. All of them have delay 10ms. So that the computed segment list would be <adj-sid@upper-link-of-S-A, adj-sid@upper-link-of-A-B, adj-sid@upper-link-of-B-C, adj-sid@upper-link-of-C-D>.

Suppose the delay of the lower member link of l2bundles interface between S-A, A-B, B-C, C-D are all 100ms. That means the delay of the bundles L3 interface between S-A, A-B, B-C, C-D are all 100ms (i.e, subject to the member who have the largest delay). Also suppose the delay of the L3 link between S-E, E-D are all 50ms.

If flex-algo (eg, algorithm 128) is enabled in the above network to optimize the stack depth of the above SR polcy, the related FAD would also be {min delay}. However, if all nodes in the network only see L3 interface resouce, then at node S the computed result to destination D would be next-hop E, and at node E the computed result to destination D would be next-hop D. Obviously, after stack optimization the flex-algo path S-E-D is not consistent with the original path (S-A-B-C-D) of SR policy.

Thus it will be benefit for flex-algo to see L2 member link during CSPF computation. And, each node in the network, instead of only headend, must perform the same behavior to check L2 member link resouce, otherwise there may be a loop.

6. IGP L2 Bundle Member Extensions

6.1. ISIS L2 Bundle Member EAG advertisement

[RFC8668] defined TLV-25 for ISIS to advertise the link attributes of L2 Bundle Members, and mentioned that the traditional "Administrative group (color) Sub-TLV" and "Extended Administrative Group Sub-TLV" may appear in TLV-25 and MAY be shared by multiple L2 Bundle Members. If we want to advertise unique EAG values for each bundle member, we can use multiple L2 Bundle Attribute Descriptors with each specify a single bundle member. So it is sufficient to construct Flex-algo plane to select L2 link resource.

6.2. OSPF L2 Bundle Member EAG advertisement

[I-D.ketant-lsr-ospf-l2bundles] defined "L2 Bundle Member Attributes sub-TLV" for OSPF/OSPFv3 to advertise the link attributes of L2 Bundle Members, and mentioned that the traditional "Administrative group (color) Sub-TLV" and "Extended Administrative Group Sub-TLV" are applicable in "L2 Bundle Member Attributes sub-TLV". Because there is "L2 Bundle Member Attributes sub-TLV" per L2 Bundle Member, it is also sufficient to construct Flex-algo plane to select L2 link resource.

6.3. FAD Flags Extensions

A new flag (L-flag) is introduced to both ISIS Flexible Algorithm Definition Flags Sub-TLV and OSPF Flexible Algorithm Definition Flags Sub-TLV (defined in [I-D.ietf-lsr-flex-algo]), to let each node to check L2 member link resouce of interface bundle during flex-algorithm path calculation.

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

Figure 2

where:

L-flag: introduced by this document. When set, the traffic engineering resouce or attributes of L2 member link of interface bundle MUST be checked and used during flex-algorithm path calculation.

7. IANA Considerations

This document need not define new sub-TLV to IGP for Flex-algo combined with l2bundles.

8. Security Considerations

There are no new security issues introduced by the extensions in this document.

9. Acknowledgements

TBD

10. Normative References

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[I-D.ketant-lsr-ospf-l2bundles]
Talaulikar, K. and P. Psenak, "Advertising L2 Bundle Member Link Attributes in OSPF", draft-ketant-lsr-ospf-l2bundles-02 (work in progress), June 2020.

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

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

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8668] Ginsberg, L., Ed., Bashandy, A., Filsfils, C., Nanduri, M., and E. Aries, "Advertising Layer 2 Bundle Member Link Attributes in IS-IS", RFC 8668, DOI 10.17487/RFC8668, December 2019, <https://www.rfc-editor.org/info/rfc8668>.

Authors' Addresses

Yongqing Zhu
China telecom

Email: zhuyq8@chinatelecom.cn

Shaofu Peng
ZTE Corporation
No.68 Zijinghua Road, Yuhuatai District
Nanjing
China

Email: peng.shaofu@zte.com.cn

Ran Chen
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China

Email: chen.ran@zte.com.cn

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Link State Routing Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 25, 2021

Y. Wang
Huawei
A. Wang
China Telecom
Z. Hu
T. Zhou
Huawei

February 21, 2021

IS-IS Multi-Flooding Instances
draft-wang-lsr-isis-mfi-00

Abstract

This document proposes a new IS-IS flooding mechanism which separates multiple flooding instances for dissemination of routing information and other types of application-specific information to minimizes the impact of non-routing information flooding on the routing convergence and stability. Due to different flooding information has different requirements on the flooding rate, these multi-flooding instances should be given various priorities and flooding parameters. An encoding format for IS-IS Multi-Flooding Instance Identifier (MFI-ID) TLV and Update Process are specified 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

Table of Contents

1. Introduction

[ISO10589] specifies the IS-IS protocol, in which each Intermediate System (IS) (router) advertises one or more IS-IS Link State Protocol Data Units (LSPs) with routing topology and Traffic Engineering (TE) information. As the one-octet LSP Number field, there are limited 256 numbers of LSPs that may be assigned. However, with the increasing amount of Topology information and TE information proposed to be advertised, for example, advertisement of Virtual Transport Networks (VTN) Topology, VTN Resource and VTN specific Data Plane Identifiers [I-D.dong-lsr-sr-enhanced-vpn], there will be huge consumption of LSPs. In addition, with the increasing use the same mechanism for advertisement of application-specific information, therefore, a mechanism should be defined for advertisement of application-specific information that minimizes the impact on the operation of the IS-IS protocol.

This document proposes a new IS-IS flooding mechanism which separates multiple flooding instances for dissemination of routing information and other types of application-specific information in a single IS-IS protocol instance. This document therefore defines an encoding format for IS-IS Multi-Flooding Instances Identifier (MFIs-ID) TLV and MFIs Update Process.

For dissemination of generic information (GENINFO) not directly related to the operation of the IS-IS protocol within the domain, [RFC6823] defines a GENINFO TLV aand specifies that the advertisement of GENINFO must occur in a non-zero instance of IS-IS protocol as defined in [RFC8202] for minimizing the impact of advertisement of GENINFO on the operation of routing. This document also recommends the use of GENINFO TLV in a specific MFI for advertisement of GENINFO in the zero IS-IS instance, which can isolate the impact of non-routing information on the standard IS-IS operation.

Instead of using non-zero IS-IS instances, the advertisement of non-routing information in MFIs is implemented in the zero IS-IS instance, which simplifies the deployment. MFIs mechanism has a lower cost to maintain neighbor because that all the MFIs share the standard IS-IS instance neighbor. In addition, MFIs can be configured with customized MFIs-specific flooding parameters (including the retransmission interval, refresh timer, maximum age, etc.).

Similarly, OSPF Multi-Flooding Instances will be proposed in the future work.

2. IS-IS Multi-Flooding Instances

An existing protocol limitation is that a given IS-IS instance in a single level supports a single update process operating on a single Link State Database (LSDB). This document defines an extension to IS-IS to allow one standard instance of the protocol to support multiple update process operations. This extension is referred to as "IS-IS Multi-Flooding Instances" (IS-IS MFIs).

Each update process is associated with a unique MFI. The behavior of the standard update process is not changed in any way by the extensions defined in this document. MFI-specific prioritization for processing PDUs and MFI-specific flooding parameters should be defined so as to allow different MFIs to consume network-wide resources at different rates. The use of MFIs can enhance the ability to isolate the resources associated with the standard update process and other application-specific update process.

2.1. Multi-Flooding Instance Identifier

A Multi-Flooding Instance Identifier (MFI-ID) is introduced to uniquely identify an IS-IS Multi-Flooding Instance and the associated update process. The protocol extension includes a new TLV (i.e. MFI-ID TLV) in each IS-IS PDU originated by an Intermediate System. It is recommended that the MFI-ID TLV be the first TLV in the PDU, which allows determination of the association of a PDU with a particular MFI more quickly. Each IS-IS PDU is associated with only one IS-IS MFI.

The MFI-ID TLV is carried in Link State PDUs (LSPs) and Sequence Number PDUs (SNPs). MFI-IDs MUST be unique within the same routing domain. The following format is used for the MFI-ID TLV:

```
                                    No. of octets
+---------------+---------------+
|      Type     |     Length    |      2
+---------------+---------------+
|            MFI-ID             |      2
+-------------------------------+
```

MFI-ID#0 is reserved for the routing flooding instance supported by legacy systems. IS-IS LSPs and SNPs do not carry the MFI-ID TLV, which indicates these PDUs are associated with the routing flooding instance in the zero IS-IS instance.

2.2. Update Process Operation

In this document, MFIs can be created in a single IS-IS instance. Different application information can be advertised to all the other Intermediate systems in the corresponding MFI.

The Update Process in an Intermediate system shall generate one or more new Link State PDUs. Each Level 1/Level 2 Link State PDU associated with a specific MFI carries application information belonging to the specific MFI. And Level 1/Level 2 PSNP and Level 1/ Level 2 CSNP containing information about LSPs that transmitted in a specific MFI are generated to synchronize the LSDB corresponding to the specific MFI.

In each MFI, update parameters can be customized differently. As specified in [ISO10589], parameters include the LSP MaxAge, LSP Refresh time, LSP retransmission interval, Maximum LSP Generation interval, Minimum LSP Generation interval, Minimum LSP transmission interval, PSNP sending interval, and CSNP sending interval. Note that besides of different update parameters, any other elements in these MFI-specific Update Process are same as the standard IS-IS

Update Process including Input and Output, Event driven LSP Generation, action on receipt of a link state PDU, etc.

2.3. Interoperability Considerations

In the scenario where some routers that do not support MFI are deployed in the same routing domain, it is recommended that all MFIs in an IS-IS protocol instance share one LSP Number space. The total number of LSPs in all MFIs cannot exceed 256. This implementation mode of MFI can coexist with routers that do not support MFI. If routers that do not support MFI receive the LSPs and SNPs encoding MFI-ID TLV, then routers SHOULD ignore the MFI-ID TLV and continues processing other TLVs.

In the scenario where all routers in the entire routing domain support MFI, it is recommended that each MFI can has its separate LSP Number space. Each MFI can have a maximum of 256 LSPs. Both LSP ID and MFI are used to uniquely identify an LSP.

Note that the MFI mechanism does not affect neighbor relationship establishment, shortest-path-first (SPF) algorithm and TE routing calculation, but only affects IS-IS LSDB synchronization.

3. IS-IS Non-routing MFIs Omission of Routing Calculation

IS-IS standard routing related TLVs and TE related extended TLVs, for example, IS Neighbors TLV and IP Reachability, are not included in Non-routing Multi-flooding Instances.

4. Applicability of IS-IS Multi-Flooding Instances

In addition to IS-IS route flooding, more and more application information and node capabilities that are not directly related to IS-IS operations need to be advertised in the entire routing domain through the IS-IS flooding mechanism. For example, the advertisement of supported In-situ Flow Information Telemetry (IFIT) capabilities at node and/or link granularity [I-D.wang-lsr-igp-extensions-ifit].

5. IANA Considerations

IANA is requested to allocate values for the following new TLV.

Type	Description
TBA	MFI-ID TLV

6. Security Considerations

It does not introduce any new security risks to IS-IS.

7. Acknowledgements

TBD

8. References

8.1. Normative References

[ISO10589]
"International Organization for Standardization, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, November 2002.", <https://www.iso.org/standard/30932.html>.

[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>.

8.2. Informative References

[I-D.dong-lsr-sr-enhanced-vpn]
"IGP Extensions for Segment Routing based Enhanced VPN", <https://datatracker.ietf.org/doc/draft-dong-lsr-sr-enhanced-vpn/>.

[I-D.wang-lsr-igp-extensions-ifit]
"IGP Extensions for In-situ Flow Information Telemetry (IFIT) Capability Advertisement", <https://datatracker.ietf.org/doc/draft-wang-lsr-igp-extensions-ifit/>.

[RFC6823] "Advertising Generic Information in IS-IS", <https://www.rfc-editor.org/info/rfc6823>.

[RFC8202] "IS-IS Multi-Instance", <https://www.rfc-editor.org/info/rfc8202>.

Authors' Addresses

Yali Wang
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: wangyali11@huawei.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
China

Email: wangaj3@chinatelecom.cn

Zhibo Hu
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: huzhibo@huawei.com

Tianran Zhou
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: zhoutianran@huawei.com

LSR Working Group A. Wang
Internet-Draft China Telecom
Intended status: Standards Track G. Mishra
Expires: September 27, 2021 Verizon Inc.
Z. Hu
Y. Xiao
Huawei Technologies
March 26, 2021

Prefix Unreachable Announcement
draft-wang-lsr-prefix-unreachable-annoucement-06

Abstract

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

Table of Contents

1. Introduction

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

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

Operators have historically relied on MPLS architecture which is based on exact match host route FEC binding for single area. [RFC5283] LDP inter-area extension provides the ability to LPM, so now the RIB match can now be a summary match and not an exact match of a host route of the egress PE for an inter-area LSP to be instantiated. SRV6 routing framework utilities the IPv6 data plane standard IGP LPM. When operators start to migrate from MPLS LSP based host route bootstrapped FEC binding, to SRv6 routing framework, the IGP LPM now comes into play with summarization which will influence the forwarding of traffic when a link or node event occurs for a component prefix within the summary range resulting in black hole routing of traffic.

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

2. Conventions used in this document

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

3. Scenario Description

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

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

```
+---------------------+------+--------+-----+--------------+
| +--+        +--+   ++-+   ++-+    +-++   + -+        +--+|
| |S1+--------+S2+---+R1+---|R0+----+R2+---+T1+--------+T2||
| +-++Ps1     +-++   ++-+   +--+    +-++   ++++    Pt2 +-++|
|   |           |      |              |      ||          | |
|   |           |      |              |      ||          | |
| +-++Ps4     +-++   ++-+           +-++   ++++     Pt4+-++|
| |S4+--------+S3+---+R3+-----------+R4+---+T3+--------+T4||
| +--+        +--+   ++-+           +-++   ++-+        +--+|
|                      |              |                    |
|                      |              |                    |
|         Area 1       |    Area 0    |     Area 2         |
+---------------------+----------------+-------------------+
```

Figure 1: OSPF Inter-Area Prefix Unreachable Announcement Scenario

3.1. Inter-Area Node Failure Scenario

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

3.2. Inter-Area Links Failure Scenario

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

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

4. PUA (Prefix Unreachable Advertisement) Procedures

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

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

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

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

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

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

In an MPLS or SR-MPLS service provider core, scalability has been a concern for operators which have split up the IGP domain into multiple areas to avoid SPF churn. Normally, MPLS FEC binding for LSP instantiation is based on egress PE exact match of a host route Looback0. [RFC5283] LDP inter-area extension provides the ability to LPM, so now the RIB match can now be a summary match and not an exact match of host route of the egress PE for an inter-area LSP to be instantiated. The caveat related to this feature that has prevented operators from using the [RFC5283] LDP inter-area extension concept is that when the component prefixes are now hidden in the summary prefix, and thus the visibility of the BGP next-hop attribute is lost.

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

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

6. Implementation Consideration

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

1. If the number of unreachable prefixes is less than MAA, the ABR should advertise the summary address and the PUA.

2. If the number of reachable address is less than MAA, the ABR should advertise the detail reachable address only.

3. If the number of reachable prefixes and unreachable prefixes exceed MAA, then advertise the summary address with MAX metric.

7. Deployment Considerations

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

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

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

8. Security Considerations

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

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

9. IANA Considerations

This document has no IANA actions.

10. Acknowledgement

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

11. Normative References

[I-D.ietf-lsr-ospf-prefix-originator]
Wang, A., Lindem, A., Dong, J., Psenak, P., and K. Talaulikar, "OSPF Prefix Originator Extensions", draft-ietf-lsr-ospf-prefix-originator-07 (work in progress), October 2020.

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

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC5283] Decraene, B., Le Roux, JL., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, DOI 10.17487/RFC5283, July 2008, <https://www.rfc-editor.org/info/rfc5283>.

[RFC5302] Li, T., Smit, H., and T. Przygienda, "Domain-Wide Prefix Distribution with Two-Level IS-IS", RFC 5302, DOI 10.17487/RFC5302, October 2008, <https://www.rfc-editor.org/info/rfc5302>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC5709] Bhatia, M., Manral, V., Fanto, M., White, R., Barnes, M., Li, T., and R. Atkinson, "OSPFv2 HMAC-SHA Cryptographic Authentication", RFC 5709, DOI 10.17487/RFC5709, October 2009, <https://www.rfc-editor.org/info/rfc5709>.

[RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <https://www.rfc-editor.org/info/rfc7770>.

[RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <https://www.rfc-editor.org/info/rfc7794>.

[RFC7981] Ginsberg, L., Previdi, S., and M. Chen, "IS-IS Extensions for Advertising Router Information", RFC 7981, DOI 10.17487/RFC7981, October 2016, <https://www.rfc-editor.org/info/rfc7981>.

Authors' Addresses

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

Email: wangaj3@chinatelecom.cn

Gyan Mishra
Verizon Inc.

Email: gyan.s.mishra@verizon.com

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

Email: huzhibo@huawei.com

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

Email: xiaoyaqun@huawei.com

LSR Working Group   Y. Zhu
Internet-Draft   China Telecom
Intended status: Standards Track   J. Dong
Expires: August 26, 2021   Z. Hu
                                                     Huawei Technologies
                                                       February 22, 2021

Using Flex-Algo for Segment Routing based VTN
draft-zhu-lsr-isis-sr-vtn-flexalgo-02

Abstract

Enhanced VPN (VPN+) aims to provide enhanced VPN service to support some application's needs of enhanced isolation and stringent performance requirements. VPN+ requires integration between the overlay VPN connectivity and the characteristics provided the underlay network. A Virtual Transport Network (VTN) is a virtual underlay network which has a customized network topology and a set of network resources allocated from the physical network. A VTN could be used as the underlay for one or a group of VPN+ services.

In some network scenarios, each VTN can be associated with a unique Flex-Algo Identifier. This document describes a mechanism to build the SR based VTNs using SR Flex-Algo and IGP L2 bundle with minor extensions.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Enhanced VPN (VPN+) is an enhancement to VPN services to support the needs of new applications, particularly including the applications that are associated with 5G services. These applications require enhanced isolation and have more stringent performance requirements than that can be provided with traditional overlay VPNs. Thus these properties require integration between the underlay and the overlay networks. [I-D.ietf-teas-enhanced-vpn] specifies the framework of enhanced VPN and describes the candidate component technologies in different network planes and layers. An enhanced VPN may be used for 5G transport network slicing, and will also be of use in other generic scenarios.

To meet the requirement of enhanced VPN services, a number of virtual transport networks (VTN) can be created, each with a subset of the underlay network topology and a set of network resources allocated from the underlay network to meet the requirement of a specific VPN+ service or a group of VPN+ services. Another possible approach is to create a set of point-to-point paths, each with a set of network resource reserved along the path, such paths are called Virtual Transport Paths (VTPs). Although using a set of dedicated VTPs can provide similar characteristics as VTN, it has some scalability issues due to the per-path state in the network.

[I-D.ietf-spring-resource-aware-segments] introduces resource awareness to Segment Routing (SR) [RFC8402]. As described in [I-D.ietf-spring-sr-for-enhanced-vpn], the resource-aware SIDs can be used to build virtual transport networks (VTNs) with the required network topology and network resource attributes to support enhanced VPN services. With segment routing based data plane, Segment Identifiers (SIDs) can be used to represent both the topology and the set of network resources allocated by network nodes to a VTN. The SIDs of each VTN and the associated topology and resource attributes need to be distributed using control plane.

[I-D.dong-lsr-sr-enhanced-vpn] defines the IGP mechanisms with necessary extensions to build a set of Segment Routing (SR) based VTNs. The VTNs could be used as the underlay of the enhanced VPN service. The mechanism described in [I-D.dong-lsr-sr-enhanced-vpn] allows flexible combination of the topology and resource attribute to build customized VTNs. In some network scenarios, each VTN can be associated with a unique Flex-Algo and allocated with a set of dedicated network resources. This document describes a mechanism to build the SR based VTNs using SR Flex-Algo and IGP L2 bundle with minor extensions.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Advertisement of SR VTN Topology Attribute

[I-D.ietf-lsr-flex-algo] specifies the mechanism to provide distributed constraint-path computation, and the usage of SR-MPLS prefix-SIDs and SRv6 locators for steering traffic along the constrained paths.

The Flex-Algo definition is the combination of calculation-type, metric-type and the topological constraints used for path computation. According to the network nodes' participation of a

Flex-Algo, and the rules of including or excluding Admin Groups (i.e. colors) and Shared Risk Link Groups (SRLGs), the topology attribute of a VTN can be described using the associated Flex-Algo. If each VTN is associated with a unique Flex-Algo, the Flex-Algo identifier could be reused as the identifier of the VTN in the control plane.

With the mechanisms defined in[RFC8667] [I-D.ietf-lsr-flex-algo], SR-MPLS prefix-SID advertisement can be associated with a specific topology and a specific algorithm, which can be a Flex-Algo. This allows the nodes to use the prefix-SIDs to steer traffic along distributed computed constraint paths according to the associated Flex-Algo in a particular topology.

[I-D.ietf-lsr-isis-srv6-extensions] specifies the IS-IS extensions to support SRv6 data plane, in which the SRv6 locators advertisement can be associated with a topology and a specific algorithm, which can be a Flex-Algo. This allows the nodes to used the SRv6 locators to steer traffic along distributed computed constraint paths according to the associated Flex-Algo in a particular topology. In addition, topology/algorithm specific SRv6 End SIDs and End.X SIDs can be used to enforce traffic over the Loop-Free Alternatives (LFA) computed backup paths.

3. Advertisement of SR VTN Resource Attribute

Each VTN can be allocated with a set of dedicated network resources. In order to perform constraint based path computation for each VTN on network controller and the ingress nodes, the resource attribute of each VTN also needs to be advertised.

[RFC8668] was defined to advertise the link attributes of the Layer-2 bundle member links. In this section, it is extended to advertise the set of network resource attributes associated with different VTNs on a shared Layer-3 link.

The Layer-3 link may or may not be a Layer-2 link bundle, as long as it has the capability of allocating different subsets of link resources to different VTNs it participates in. A subset of the link resources can be considered as a virtual Layer-2 member link (or sub-interface) of the Layer-3 link. If the Layer-3 interface is a Layer-2 link bundle, it is possible that the subset of link resource allocated to a specific VTN is provided by one of the physical Layer-2 member links.

A new flag "V" (Virtual) is defined in the flag field of the Parent L3 Neighbor Descriptor in the L2 Bundle Member Attributes TLV (25).

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

V flag: When the V flag is set, it indicates the advertised member links under the Parent Layer-3 link are virtual Layer-2 member links. When the V flag is clear, it indicates the member links are physical member links. This flag may be used to determine whether the member links share fates with the parent interface.

For each virtual or physical member link, the TE attributes defined in [RFC5305] such as the Maximum Link Bandwidth and Admin Groups SHOULD be advertised using the mechanism as defined in [RFC8668]. The Adj-SIDs or SRv6 End.X SIDs associated with each of the virtual or physical Layer-2 member links SHOULD also be advertised.

In order to correlate the virtual or physical member links with the Flex-Algo used to identify the VTN, each VTN SHOULD be assigned with a unique Admin Group (AG) or Extended Admin Group (EAG), and the virtual or physical member link associated with this VTN SHOULD be configured with the AG or EAG assigned to the VTN. The AG or EAG of the Layer 3 link SHOULD be set to the union of all the AGs or EAGs of its virtual or physical member links. In the definition of the Flex-Algo corresponding to the VTN, It MUST use the Include-Any Admin Group rule with only the AG or EAG assigned to the VTN as the link constraints, the Include-All Admin Goup rule or the Exclude Admin Group rule MUST NOT be used. This ensures that the Layer-3 link is included in the Flex-Algo specific constraint path computation for each VTN it participates in.

4. Forwarding Plane Operations

For SR-MPLS data plane, a prefix SID is associated with the paths calculated using the corresponding Flex-Algo of a VTN. An outgoing Layer-3 interface is determined for each path. In addition, the prefix-SID also steers the traffic to use the virtual or physical member link which is associated with the VTN on the outgoing Layer-3 interface for packet forwarding. The Adj-SIDs associated with the virtual or physical member links of a VTN MAY be used with the prefix-SIDs of the same VTN together to build SR-MPLS paths with the topological and resource constraints of the VTN.

For SRv6 data plane, an SRv6 Locator is a prefix which is associated with the paths calculated using the corresponding Flex-Algo of a VTN. An outgoing Layer-3 interface is determined for each path. In addition, the SRv6 Locator prefix also steers the traffic to use the virtual or physical member link which is associated with the VTN on the outgoing Layer-3 interface for packet forwarding. The End.X SIDs associated with the virtual or physical member links of a VTN MAY be used with the SRv6 Locator prefix of the same VTN together to build SRv6 paths with the topological and resource constraints of the VTN.

5. Scalability Considerations

The mechanism described in this document assumes that each VTN is associated with an unique Flex-Algo, so that the Flex-Algo IDs can be reused to identify the VTNs in the control plane. While this brings the benefit of simplicity, it also has some limitations. For example, it means that even if multiple VTNs share the same topological constraints, they would still need to be identified using different Flex-Algo IDs in the control plane, then independent path computation needs to be executed for each VTN. The number of VTNs supported in a network may be dependent on the number of Flex-Algos supported, which is related to the control plane computation overhead. Another aspect which may impact the number of VTNs supported with this mechanism is that at most 128 Flex-Algos can be used in a network.

Based on the above considerations, this mechanism may be suitable for networks where a relatively small number of VTNs are needed.

6. Security Considerations

This document introduces no additional security vulnerabilities to IS-IS.

The mechanism proposed in this document is subject to the same vulnerabilities as any other protocol that relies on IGPs.

7. IANA Considerations

This document does not request any IANA actions.

8. Acknowledgments

The authors would like to thank Zhenbin Li and Peter Psenak for the review and discussion of this document.

9. References

9.1. Normative References

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extension to Support Segment Routing over IPv6 Dataplane", draft-ietf-lsr-isis-srv6-extensions-11 (work in progress), October 2020.

[I-D.ietf-spring-resource-aware-segments]
Dong, J., Bryant, S., Miyasaka, T., Zhu, Y., Qin, F., Li, Z., and F. Clad, "Introducing Resource Awareness to SR Segments", draft-ietf-spring-resource-aware-segments-01 (work in progress), January 2021.

[I-D.ietf-spring-sr-for-enhanced-vpn]
Dong, J., Bryant, S., Miyasaka, T., Zhu, Y., Qin, F., Li, Z., and F. Clad, "Segment Routing based Virtual Transport Network (VTN) for Enhanced VPN", February 2021, <https://tools.ietf.org/html/draft-ietf-spring-sr-for-enhanced-vpn>.

[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>.

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

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC8668] Ginsberg, L., Ed., Bashandy, A., Filsfils, C., Nanduri, M., and E. Aries, "Advertising Layer 2 Bundle Member Link Attributes in IS-IS", RFC 8668, DOI 10.17487/RFC8668, December 2019, <https://www.rfc-editor.org/info/rfc8668>.

9.2. Informative References

[I-D.dong-lsr-sr-enhanced-vpn]
Dong, J., Hu, Z., Li, Z., Tang, X., Pang, R., JooHeon, L., and S. Bryant, "IGP Extensions for Segment Routing based Enhanced VPN", draft-dong-lsr-sr-enhanced-vpn-04 (work in progress), June 2020.

[I-D.ietf-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-ietf-spring-srv6-network-programming-28 (work in progress), December 2020.

[I-D.ietf-teas-enhanced-vpn]
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Networks (VPN+) Service", draft-ietf-teas-enhanced-vpn-06 (work in progress), July 2020.

Authors' Addresses

Yongqing Zhu
China Telecom

Email: zhuyq8@chinatelecom.cn

Jie Dong
Huawei Technologies

Email: jie.dong@huawei.com

Zhibo Hu
Huawei Technologies

Email: huzhibo@huawei.com