IS-IS for IP Internets
Internet-Draft
Intended status: Standards Track
Expires: September 21, 2013

S. Previdi, Ed.
C. Filsfils, Ed.
A. Bashandy
Cisco Systems, Inc.
M. Horneffer
Deutsche Telekom
B. Decraene
S. Litkowski
Orange
I. Milojevic
Telekom Srbija
R. Shakir
British Telecom
S. Ytti
TDC Oy
W. Henderickx
Alcatel-Lucent
J. Tantsura
Ericsson

March 20, 2013

Segment Routing with IS-IS Routing Protocol
draft-previdi-filsfils-isis-segment-routing-02

Abstract

Segment Routing (SR) enables any node to select any path (explicit or derived from IGPs SPT computations) for each of its traffic classes. The path does not depend on a hop-by-hop signaling technique (neither LDP nor RSVP). It only depends on a set of "segments" that are advertised by the IS-IS routing protocol. These segments act as topological sub-paths that can be combined together to form the desired path.

There are two forms of segments: node and adjacency. A node segment represents a path to a node. An adjacency segment represents a specific adjacency to a node. A node segment is typically a multi-hop path while an adjacency segment is a one-hop path. SR's control-plane can be applied to IPv6 and MPLS dataplanes.

Segment Routing control-plane can be applied to the MPLS dataplane: a node segment to node N is instantiated in the MPLS dataplane as an LSP along the shortest-path (SPT) to the node. An adjacency segment is instantiated in the MPLS dataplane as a cross-connect entry pointing to a specific egress datalink.

This document describes the Segment Routing functions, a set of use cases it addresses and the necessary changes that are required in the IS-IS protocol.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on September 22, 2013.

Copyright Notice

Table of Contents

1. Introduction

Segment Routing (SR) enables any node to select any path (explicit or derived from IGPs SPT computations) for each of its traffic classes. The path does not depend on a hop-by-hop signaling technique (neither LDP nor RSVP). It only depends on a set of "segments" that are advertised by the IS-IS routing protocol. These segments act as topological sub-paths that can be combined together to form the desired path.

There are two forms of segments: node and adjacency. A Node Segment represents the shortest path to a node. A Node Segment is typically a multi-hop sortest path. An adjacency Segment represents a specific adjacency to a node.

SR's control-plane can be applied to IPv6 and MPLS dataplanes.

In the MPLS dataplane, a node segment to node N is instantiated as an LSP along the shortest-path (spt) to the node. An adjacency segment

is instantiated as a crossconnect entry pointing to a specific egress datalink.

At the heart of the SR technology, we find node segments. Node segments must be globally unique within the network domain.

Figure 1

In Figure 1, all the nodes must be configured with the same Segment Routing Identifiers Block (called SRB Node Registry), e.g. 64-5000, and any node segment be uniquely allocated from that SRB Node Registry (e.g. A, B, C and D are configured respectively with node segments 64, 65, 66 and 67).

In the MPLS dataplane instantiation, this means that all the nodes need to be able to reserve and allocate to the SR control-plane the same MPLS label range (e.g. 64-5000).

2. Applicability

Segment Routing is applicable to the following use-cases: simplicity, TE, FRR and SDN.

2.1. Simplicity

The vast majority of IP traffic travels on shortest-paths to their destination. SR delivers a very efficient control-plane technique to instantiate shortest-path-based node segments into the forwarding dataplane. In the example described in Figure 1, considering the MPLS forwarding plane, when node D advertises node segment 64 for its loopbacks D/32, node A and B introduce the following MPLS Dataplane entries:

A: IP2MPLS: FEC D/32 => push 64, nhop B
A: MPLS2MPLS: 64 => swap 64, nhop B
B: IP2MPLS: FEC D/32 => push 64, nhop C
B: MPLS2MPLS: 64 => swap 64, nhop C

If D advertises node segment 64 with the P flag reset:

C: IP2MPLS: FEC D/32 => push explicit-null, nhop D
C: MPLS2MPLS: 64 => pop, nhop D

If D advertises node segment 64 with the P flag set:

```
C: IP2MPLS: FEC D/32 => push 64, nhop D
C: MPLS2MPLS: 64 => swap 64, nhop D
```

LDP is no longer required to instantiate shortest-path LSP's to a remote node. The reduction in the number of protocols to operate, helps reduce the overall operational complexity of the network. For example, the complex IGP/LDP synchronization, described in [RFC5443] and [RFC6138] no longer needs to be considered hence drastically improving the scaling and reliability of the network.

For example, when a core node C has 40 TE tunnels to 40 remote core routers and 260 adjacent aggregation routers and LDP LSP's need to be signaled to 5000 FEC's, then node C maintains an LDP label database of (260+40)*5000 = 1.500.000 label bindings. In fact several networks today are exposed to much more difficult LDP scaling constraints.

In comparison, in the same use case, SR control-plane only maintains 5000 node segments. This is 300 times more scalable.

2.2. Capacity Planning and Traffic Engineering (TE)

Capacity Planning deals with anticipating the placement of the traffic matrix onto the network topology, for a set of expected traffic and topology variations. The heart of the process consists in simulating the placement of the traffic along ECMP-aware shortest-path and accounting for the resulting bandwidth usage. The bandwidth accounting of a demand along its shortest-path is a basic capability of any planning tool or PCE server.

For example, in the network topology described in Figure 2 and assuming a default IGP metric of 2 and IGP metrics BC=BG=CD=CE=DF=EF=1, a 1600Mbps A-to-Z flow is accounted as consuming 1600Mbps on links AB and FZ, 800Mbps on links BC, BG and GF, and 400Mbps on links CD, DF, CE and EF.

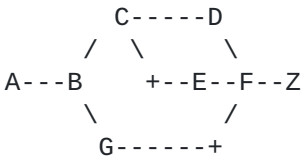


Figure 2: Example Topology 1

ECMP is extremely frequent in SP, Enterprise and DC architectures and it is not rare to see as much as 128 different ECMP paths between a source and a destination within a single network domain.

This is illustrated in Figure 3 which consists of a subset of a network where already 6 ECMP paths are observed from A to M.

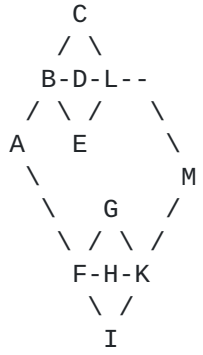


Figure 3: ECMP Topology Example

Segment Routing offers a simple support for such ECMP-based shortest-path placement: a node segment. A single node segment enumerates all the ECMP paths along the shortest-path.

This is much simpler to the RSVP-TE model where one TE tunnel is required for each enumerated ECMP path.

When the capacity planning process detects that a traffic or topology variation would lead to congestion traffic engineering or capacity increase is triggered.

The most basic traffic engineering option consists of finding the smallest set of demands that need to be routed off their shortest path to eliminate the congestion, then to compute an explicit path for each of them and instantiating these traffic-engineered policies in the network.

Segment Routing offers a simple support for explicit path policy.

In the diagram described in Figure 3, it is assumed that the requirement is that AM flow should not consume any resource on the LM and the FG links.

The first option would consists of using the following encapsulation at A: A sends any traffic to M towards the nhop F with a two-label stack where the top label is the adjacent segment FI and the next label is the node segment to M. Alternatively, a three-label stack with adjacency segments FI, IK and KM could have been used.

The first option seems preferred as classically IP capacity planners optimize traffic along ECMP-aware shortest-path. The more node

segment can be used, the better. However, both options are available and one can favor adjacency segments.

In the same way, if the requirement in the diagram described in Figure 3, is that the AM flow should not consume any resource along the LM link but should use any resource on the bottom of the topology, then A could send the AM flow to its nhop F with a single label: the label representing the node segment to M.

We believe that Segment Routing offers an excellent solution for Capacity Planning because:

Node segment represents an ECMP-aware shortest path.

Adjacency segments allow to express any explicit path.

The combination of node and adjacency segment allows to express any path without having to enumerate all the ECMP path options.

The capacity planning process ensures that the majority of the traffic rides on node segments (ECMP-based shortest path) while a minority of the traffic is routed off its shortest-path.

The network does not hold any signaling state for the traffic engineered flows.

In comparison, a classic RSVP-TE Full-mesh traffic engineering solution involves a full-mesh of tunnels from any edge to edge of the network. For any specific edge to edge pair, tens of RSVP-TE tunnels may need to be enumerated to load-share the traffic along all the possible shortest paths.

Analytically, assuming a single tunnel from an edge to an edge (optimistic assumption), an RSVP-TE Full-Mesh traffic engineering solution scales as E^2 where E is the number of edge nodes. The number of LSP's signaled and maintained by the network (in control-plane and in dataplane) scales quadratically with the number of edge nodes.

In contrast, the Segment Routing solutions maintains E node segments. The number of control-plane and dataplane states scale linearly with the number of edge nodes.

A network of 1000 edges is very frequent nowadays. In such a case, the capacity planning solution based on segment routing scales 1000 times better than the RSVP-TE Full-Mesh solution.

We have applied this comparative study to a use-case using real topology and real demand matrix. The data-set consisted in a full-mesh of 12000 Tunnels where originally only 65% of the traffic was riding on their shortest path. Two well-known defects are illustrated in this data set: the lack of ECMP support in RSVP-TE and hence the increase of the number of tunnels to enumerate all the ECMP options, the inefficiency of distributed optimization as too much traffic is riding off its shortest path. Using centralized optimization, we could optimize the IGP metrics such as to place 98% of the traffic on ECMP-aware shortest-path (one single node segment) while only 2% of the traffic required explicit traffic engineering tunnels away from the shortest path. Only 250 demands required explicit paths.

In this example, we increased the efficiency of the network by 150%. Indeed, 98% is riding on shortest path instead of 65%. Furthermore, we reduced the operational complexity of the network by 60 times (200 explicit routing policies instead of 12000).

The next two sections provide other examples illustrating the simplicity and efficiency benefits of the SR-based traffic engineering solution.

2.2.1. Disjointness in dual-plane networks

Many networks are built according to the dual-plane design:

Each access region k is connected to the core by two C routers (C(1,k) and C(2,k)).

C(1,k) is part of plane-1 of the dual-plane core.

C(2,k) is part of plane-2 of the dual-plane core.

C(1,k) has a link to C(2, l) iff k = l.

{C(1,k) has a link to C(1, l)} iff {C(2,k) has a link to C(2, l)}.

Many networks need to deliver disjoint-based services (bank, government...): an access node A connected to core nodes C(1, A) and C(2, A) need to provide two disjoint services towards an access node Z connected to core nodes C(1, Z) and C(2, Z).

Classic IP routing cannot fulfill this requirement as A would load-balance between the dual planes across ECMP paths.

RSVP-TE traffic-engineering would allow to signal two disjoint paths: one across the first plane and one across the second plane with the following two draw-backs:

Many ECMP paths exist within each plane (from (Ci, A) to (Ci, Z)) and hence many RSVP-TE tunnels might be required to efficiently distribute the load within each plane.

Many such services might need to be supported.

Assuming 10000 such services across the network and assuming an average of 4 ECMP paths within each plane, a straight application of RSVP-TE would require 10000 * 2 * 4 tunnels hence 80000 tunnels. Even if load-sharing of traffic along ECMP paths in each plane is dropped, the solution would still need 20000 tunnels.

Segment Routing (SR) offers a simpler solution.

Any node of the first plane can be configured with an anycast loopback say 1.1.1.1/32 to which node segment 111 is attached. Any node of the second plane can be configured with an anycast loopback say 2.2.2.2/32 to which node segment 222 is attached. Let us also assume that access node Z is advertising node segment 500.

A flow from A to Z via the first plane is simply represented by the segment list {111, 500}. In the MPLS dataplane case, A pushes a two-label stack for Z-destined packets: the top label is 111 and the second label is 500.

Node segment 111 gets the traffic on ECMP-aware shortest path to the first plane and then node segment 500 gets the traffic on ECMP-aware shortest path to Z.

Similarly, a flow from A to Z via the second plane is simply represented by the segment list {222, 500}.

This simple solution would only add two node segments to the network instead of 80000 LSP's signaled by the RSVP-TE solution. This is 40000 better.

2.2.2. QoS-based Routing Policies

Frequently, different classes of service need different path characteristics.

For example, an international network with presence in Tokyo and Brussels may have lots of cheap network capacity from Tokyo to Europe via USA and some scarce expensive capacity via Russia.

...USA...Brussels...Russia...Tokyo...USA...

Figure 4: International Topology Example

In such case, the IGP metrics would be tuned to have a shortest-path from Tokyo to Brussels via USA.

This would provide efficient capacity planning usage while fulfilling the requirements of most of the data traffic. However, it may not suite the latency requirements of the voice traffic between the two cities.

Segment Routing (SR) offers a simple solution to the problem.

The core routers in Russia are configured with an extra anycast loopback 3.3.3.3/32 to which node segment 333 is attached.

If we assume that Brussels is configured with node segment 600, Tokyo can send all its data traffic to Brussels with one single segment: 600. 600 gets the traffic from Tokyo to Brussels via USA and exploits any ECMP-path along this shortest-path.

Tokyo can send all its voice traffic to Brussels with a list of two segments: {333, 600}. 333 gets the traffic to Russia and exploit any ECMP path along the shortest path. 600 gets the traffic from Russia to Brussels via ECMP-aware shortest-path.

One single metric per link is sufficient as clearly it is possible to set the IGP metrics such that the shortest-path from Brussels to Russia is via Russia and not via USA and the shortest-path from Russia to Brussels is not back via Tokyo and USA but straight to Brussels.

2.2.3. Deterministic non-ECMP Path

The previous sections have illustrated the ease of capacity planning traffic with ECMP-awareness and shortest-path. The key benefits in terms of drastic reduction of the number of routing policies signaled by the network control plane and maintained by the data plane have been explained and several orders of scaling simplifications have been illustrated.

In this section, we highlight SR's ability to support a completely different model: the deterministic expression of a path avoiding any ECMP behavior. This is realized by expressing the end-to-end path as a list of adjacency segments.

For example, in Figure 3, one can force the AM traffic on the explicit path AFGKM by using the segment list {AF, FG, GK, KM}.

Once again, SR offers simplicity and scaling benefits compared to the alternative RSVP-TE solution: no state is signaled through the network.

In Figure 3, with SR, nodes F, G, K and M do not maintain any SR state for the A-to-M policy. With RSVP-TE, each nodes along the RSVP-TE tunnel must maintain one LSP state per tunnel.

Here is a technique to decrease the number of adjacency segments to describe non-ECMP paths.

In the topology example illustrated in Figure 5 node C can be configured with an SR explicit policy to node G via the path CDEFG.

A-B-C-D-E-F-G-H

Figure 5: Topology Example 3

Node C can advertise a (forwarding) adjacency to node G and attach an SR subTLV to identify the related adjacency segment (e.g 72). The ERO SubTLV is further attached to identify that this adjacency is not describing a real datalink between C and G but instead an SR non-ECMP sub-path along the route {BC, CD, DE, EF, FG}.

Node A can then express its desired non-ECMP path has {AB, BC, 72, GH} instead of {AB, BC, CD, DE, EF, FG, GH}.

Future versions of the document will document other techniques to decrease the number of adjacency segments in non-ECMP source-routed paths.

2.3. Fast Reroute

This section assumes familiarity with Remote-LFA concepts described in [I-D.ietf-rtgwg-remote-lfa].

Lemma 1: In networks with symmetric IGP metrics (the metric of a link AB is the same as metric of the reverse link BA), we can prove that either the P and the Q sets intersect or there is at least one P node that is adjacent to a Q node.

Consider an arbitrary protected link S-E. In LFA FRR, if a path to the destination from a neighbor N of S does not cause a packet to loop back over the link S-E (i.e. N is a loop-free alternate), then S can send the packet to N and the packet will be delivered to the destination using the pre-failure forwarding information.

If there is no such LFA neighbor, then S may be able to create a virtual LFA by using a tunnel to carry the packet to a point in the network which is not a direct neighbor of S and from which the packet will be delivered to the destination without looping back to S. Remote LFA (RLFA, [I-D.ietf-rtgwg-remote-lfa]) calls such a tunnel a repair tunnel. The tail-end of this tunnel is called a "remote LFA" or a "PQ node". We refer to RLFA for the definitions of the P and Q sets.

If there is no such RLFA PQ node, we propose to use a Directed LFA (DLFA) repair tunnel to a Q node that is adjacent to the P space. The

DLFA repair tunnel only requires two segments: a node segment to a P node which is adjacent to the Q node and an adjacency segment from the P node to its adjacent Q node.

It results from lemma1, that thanks to the DLFA extension, we have a guaranteed LFA-based FRR technique for any network with symmetric IGP metrics.

The solution is simple:

- it does not require any extra computation on top of the one required for RLFA.

- The repair tunnel can be encoded efficiently with only two segments.

The solution preserves the capacity planning properties of LFA FRR.

2.4. Segment Routing in Software Defined Networks (SR-SDN)

Some of the SDN requirements are:

- Guarantees of Tight SLA's (FRR and bandwidth admission control).

- Efficient use of the network resources.

- Very high scaling to support application-based transactions.

Segment Routing (SR) is a compelling architecture to support SDN for the following reasons.

SR supports a simple but efficient capacity planning process based on centralized optimization.

SR optimizes network resources by providing a very simple support for ECMP-based shortest-path flows

SR provides for much better scaling than alternative solution: several orders of scaling gains have been illustrated in the simplicity and Capacity Planning sections.

SR provides for guaranteed-FRR for any topology.

SR provides for ultimate virtualization as the network does not contain any application state. The state is in the packet. It is encoded as a list of segments.

SR provides for very frequent transaction-based application as the network does not hold any state for the SR-encoded flows.

3. Segment Routing Identifiers

Segment Routing defines two types of Segment Identifiers: Node-SID and Adj-SID.

3.1. Node Segment Identifier (Node-SID)

A node-SID is associated to a prefix advertised by a node (e.g. in a TLV-135). The Node-SID SubTLV MAY be present in one of the following TLVs:

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

TLV-235 (MT-ipv4) [RFC5120].

TLV-236 (IPv6) [RFC5308].

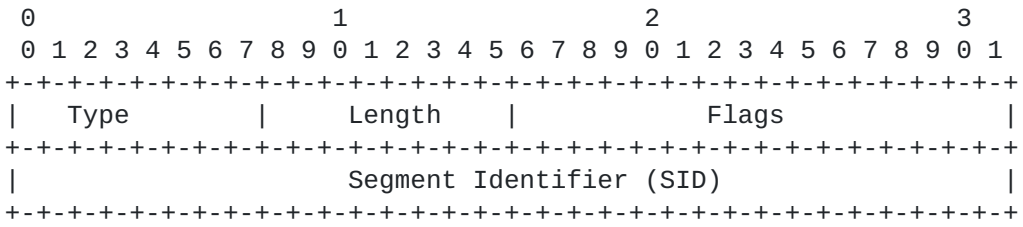
TLV-237 (MT-IPv6) [RFC5120].

Multiple Node-SID SubTLVs MAY be attached to a prefix. A node receiving a Node-SID subTLV containing more than one Node-SID MAY consider only one encoded Node-SID, in which case, the first encoded Node-SID MUST be considered and any additional Node-SID ignored.

The value of the Node-SID is a 32 bit number.

3.1.1. Node-SID SubTLV

The Node-SID SubTLV has the following format:

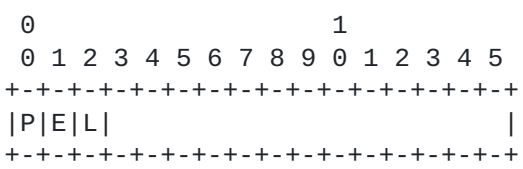


where:

Type: TBA

Length: 6 octets

Flags: 2 octets field of following flags:



where:

P-Flag: PHP flag. If set, then the penultimate hop MUST NOT pop the Nodal-SID before delivering the packet to the node that advertised the Node-SID.

E-Flag: External flag. If set, then the prefix is not local to the ISIS protocol. It is redistributed from another protocol.

L-Flag: Level flag. If set, then the prefix has been propagated to the router in this level from another level (i.e.: from level-1 into level-2 or from level-2 into level-1).

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

Segment Identifier (SID): 32 bits of Segment Identifier

3.2. Adjacency Segment Identifier (Adj-SID)

An Adjacency Segment Identifier (Adj-SID) represents a router adjacency. The value of the Adj-SID is local to the router and it is encoded as a 32 bit number value using a new SubTLV. According to IS-IS, each adjacency is advertised using one of the IS-IS Neighbor TLVs below:

TLV-22 [RFC5305]

TLV-222 [RFC5120]

TLV-23 [RFC5311]

TLV-223 [RFC5311]

TLV-141[RFC5316]

Currently, [RFC5316] defines TLV-141 with the purpose of inter-AS connectivity. In the Segment Routing context, we relax the constraint and we allow TLV-141 to be used for advertising any link that is external to the IS-IS domain no matter if it connects another AS or not.

The newly defined Adj-SID subTLV carries the Adj-SID value for each of the advertised adjacencies and MAY be present in any of the neighbor TLVs described above.

Multiple Adj-SID SubTLVs MAY be attached to the Neighbor TLVs (e.g.: TLV-22). An example where more than one is useful is the case of parallel adjacencies between two neighbors. In the figure here below:

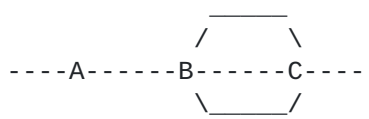


Figure 6: Parallel Adjacencies

Router B nd C have 3 parallel adjacencies. Router B advertises three distinct Neighbor TLVs (e.g.: TLV-22), one for each parallel adjacency. Each of these advertisements will have its own Adj-SID SubTLV with a unique value (inside the Adj-SID space of the router).

When router A inspects its IS-IS Link State Database (LSDB) it can figure out which link to use on a source routed path going through B-C links. It has knowledge of each individual parallel adjacency and can handle load sharing across them on its own (i.e.: decide in advance which packet should use which link).

However, router A may prefer not to select a specific parallel interface and leave the load sharing decision to router B so that load sharing is handled locally (i.e.: where parallel interfaces resides).

In order to achieve that, router B inserts an additional Adj-SID value on each of the parallel adjacencies it advertises. The value of this second Adj-SID is common to all parallel adjacencies.

Again, when router A inspects its IS-IS LSDB, it finds that the parallel adjacencies advertised by router B have a second Adj-SID with a value that is common across all parallel adjacencies. Using that value will bring packets into router B and the load sharing decision is owned by router B itself.

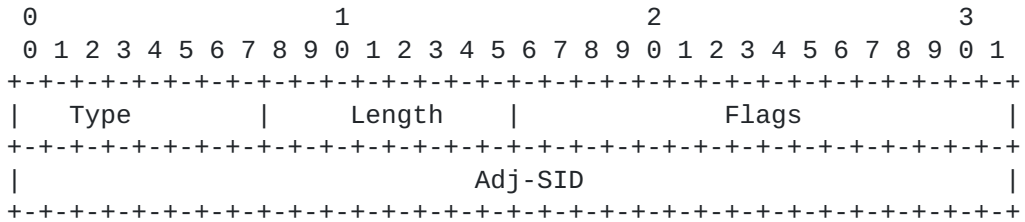
When the same Adj-SID value is used on parallel adjacencies, we called the Adj-SID a "Bundle-Adj-SID".

3.2.1. Adj-SID and Interface Address

When advertising one or more Adj-SID SubTLVs, the router MUST also advertise Interface Address and Neighbor Address SubTLVs (IPv4 or IPv6). The two MUST be present. The encoding is defined in [RFC5305] for IPv4 and in [RFC6119] for IPv6.

3.2.2. Adjacency Segment Identifier (Adj-SID) SubTLV

The following format is defined for the Adj-SID.

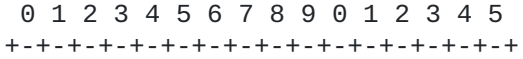


where:

Type: TBA

Length: variable.

Flags: 2 octets field of following flags:



```
|B|F|                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

B-Flag: Bundle flag. If set, then Adj-SID refers to a bundle (i.e.: a set of parallel adjacencies).

F-Flag: FA flag. If set, then Adj-SID refers to a Forwarding Adjacency.

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

Adj-SID: 32 bits of Adjacency Segment Identifier

Forwarding Adjacencies are defined in [RFC4206].

If the F-flag is set, then the explicit path taken by the Forwarding Adjacency MUST be encoded using the following subTLV in the Adj-SID SubTLV:

```
 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             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Segment Identifier (SID) #1               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Segment Identifier (SID) #...             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

Type: TBA.

Flags: none are currently used.

Length: variable, 2 + multiple of 4 octets.

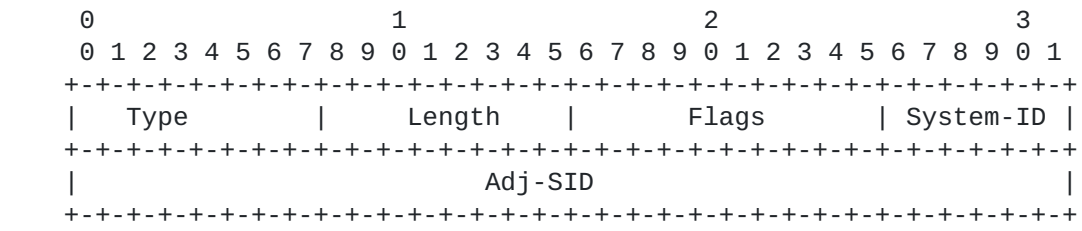
Segment Identifier (SID): The SID value of each hop in the explicit path of the Forwarding Adjacency.

3.2.3. Adjacency Segment Identifiers in LANs

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

Still, when Segment Routing is used, each router in the LAN MUST advertise the Adj-SID of each of its neighbors. Since, on LANs, there are no neighbor advertisements in non-PN-LSPs (other than the adjacency to the DIS), each router advertises the set of Adj-SIDs (for each its neighbors) inside the Intermediate To Intermediate Hello (IIH) packets as soon as the adjacency to that neighbor reaches the UP state.

We define a new IIH TLV, the IIH-Adj-SID TLV with following format:



Where:

Type: TBA

Length: 6 octets

Flags: 10 bits of flags. None are used at this stage. MUST be zero when sent and ignored when received.

System-ID: 6 octets of system ID and pseudonode number of the neighbor.

Adj-SID: 32 bits of IIH Adjacency Segment Identifier

Therefore, each router in the LAN advertises in its IIH packet the list of UP adjacencies in the form of tuples: <SystemID, Adj-SID>.

The DIS, as any other router in the LAN, receives IIHs from all routers on the LANand stores the set of tuples <System-ID, Adj-SID>.

The DIS includes the Adj-SID information received in the IIHs when advertising IS-Neighbors in its PN-LSPs.

The result is that the PN-LS contains the neighbors of the DIS and, for each of them, the list of their Adj-SIDs to their respective neighbors in the LAN.

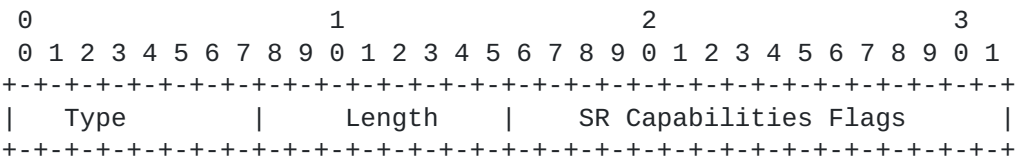
This could require multiple IS-Neighbor TLVs for the same neighbor if there are more than 25 ISs on a LAN.

Each router within the level-1 area or level-2 subdomain, when receiving the PN-LSP, will extract each neighbor and its corresponding Adj-SID table in order to figure out which Adj-SID has to be used between any two neighbors in the LAN.

4. Segment Routing Capabilities

Segment Routing requires each router to advertise its capabilities to the rest of the routing domain. TLV-242 (defined in [RFC4971]) describes router capabilities. For the purposes of Segment Routing we define an additional subTLV: the SR-Cap SubTLV.

The SR-Cap SubTLV MUST be present in the Router Capability TLV (TLV-242), MUST appear only once and has following format:

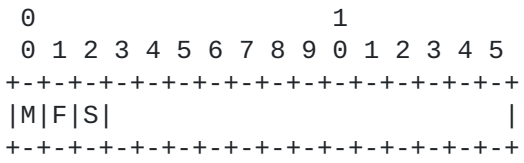


where:

Type: TBA.

Length: 2 octets.

SR Capabilities Flags: 2 octets field of following flags:



where:

M-Flag: MPLS flag. If set, then the advertising router is capable of MPLS label based forwarding.

F-Flag: IPv4 flag. If set, then the advertising router is capable of IPv4 based forwarding.

S-Flag: IPv6 flag. If set, then the advertising router is capable of IPv6 based forwarding.

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

The Router Capability TLV defined in [RFC4971]specifies the S and D bits. The SR-Capability SubTLV MUST be propagated throughout the entire routing domain and therefore the S bit in the Router Capability TLV MUST be set.

The D bit of Router Capability TLV must be set accordingly. I.e.: it MUST be set when the Router Capability TLV is leaked from level-2 to level-1.

5. Elements of Procedure

This section describes aspects of Segment Routing procedures.

5.1. Unicity

The benefits of the Segment Routing solution build up on a small set of rules. The first 64 values of the 32-bit segment space are reserved and cannot be used by the SR Control-Plane neither for node or adjacency segment.

All the nodes in the ISIS domain must be configured with the node SRB range. The range is a local policy and is not advertised by ISIS. A node segment must be allocated from the node SRB range.

A given Node-SID must be allocated to a unique IP prefix. If the IP prefix is of anycast type and is advertised by two nodes N and M, then N and M attach the same (anycast) Node-SID to the same anycast IP address.

If a node N learns a remote Adj-SID S but advertised with a value that falls in its locally configured Node SRB range, N SHOULD issue an error log warning for a misconfiguration.

If a node N learns a remote Node-SID S but with a value that falls outside its locally configured node SRB range, N SHOULD NOT insert any RIB entry for segment S. Node N SHOULD issue an error log warning for misconfiguration.

If a node N learns about two different IP addresses advertised with the same Node-SID, N MUST insert a RIB entry only for the node segment related to the highest IP address. N SHOULD issue an error log warning for misconfiguration.

5.2. IS-IS Multi-Level

In IS-IS protocol, adjacencies advertisements (e.g.: TLV-22) are not propagated across level/area boundaries hence the adjacency segment (Adj-SID) is not propagated across levels either.

If a prefix is propagated across levels, then its Node-SID SubTLVs are also propagated. The Node-SID S flag is set accordingly, independently from the settings of the U/D bit defined in [RFC5305].

5.3. Data-Plane Encodings

The SR control-plane supports different forwarding planes. The first section describes the SR source routing concept and its RIB representation. The next sections map the SR-RIB entries into the MPLS and IPv6 forwarding planes.

5.3.1. Segment Routing RIB (SR-RIB)

SR leverages source routing and introduces the following terminology:

A packet is prepended with an SR header which contains a list of segments.

A list of segments is ordered and has a pointer identifying the active segment.

The active segment is the segment identified by the pointer.

Forwarding is based on the active segment.

The following forwarding operations are defined for SR:

CONTINUE: the active segment remains active after the forwarding operation and the pointer is left unchanged.

NEXT: the active segment is completed after the forwarding operation and the pointer is advanced to the next segment in the ordered list.

INSERT: a list of segments is inserted in the segment list. The INSERT operation can be coupled with the CONTINUE or NEXT operation.

Other operations will be introduced in future versions of the document.

Two types of SR-RIB entries are defined:

TRANSIT: the ingress packet comes with an active segment. A Transit SR-RIB entry is represented as:

Ingress active segment.

Operation on the active segment.

Egress Interface.

INGRESS: the ingress packet comes without active segment (plain IP).

5.3.1.1. SR-RIB entry for local segments

A node MUST install a transit SR-RIB entry for any local adjacency segment (Adj-SID) of value V attached to datalink L with:

Ingress active segment : V

Ingress operation: NEXT

Egress interface: L

A node MUST install a transit SR-RIB entry for any local adjacency segment (Adj-SID) of value W attached to ISIS link bundle B with:

Ingress active segment: W

Ingress operation: NEXT

Egress interface: hash between any datalink within bundle B

A node MUST install a transit SR-RIB entry for any local node segment (Node-SID) of value N with:

Ingress active segment: N

Ingress operation: NEXT (if not the last segment, then process the next segment else lookup in IP table)

5.3.1.2. Transit SR-RIB entry for remote segments

A node MUST install a transit SR-RIB entry for any remote node segment (Node-SID) of value R attached to IP prefix P with:

Ingress active segment: R

Ingress operation: CONTINUE (However, if the P flag is reset and P is advertised by the next-hop, then the operation is NEXT instead of CONTINUE).

Egress interface: interface to next-hop along the shortest-path to P.

A transit SR-RIB entry is never installed for a remote adjacency segment.

5.3.1.3. Ingress SR-RIB entry for remote segments

Ingress SR-RIB entries enable traffic injection in the SR forwarding plane. An ingress SR-RIB entry is generally represented as:

Classification: what traffic

Encapsulation: what list of segments to insert

In this section, we define its simplest instantiation: the automated ingress SR-RIB entry insertion towards remote node segments (Node-SID).

A node SHOULD install an ingress SR-RIB entry for any remote node segment (Node-SID) of value V attached to IP prefix P with:

FEC: prefix P

Ingress operation: insert nodal segment V.

Egress interface: interface to next-hop along the shortest-path to P.

5.3.1.4. Policy-based Ingress SRIB entry

The text will be added in future revision.

5.3.2. Multiprotocol Label Switching (MPLS)

The mapping of SR-RIB entries into the MPLS forwarding plane is straightforward. The following elements MUST be considered:

A list of segments is represented as a stack of labels.

The active segment is the top label.

The CONTINUE operation is implemented as a swap where the outgoing label value is set to the incoming label value.

The NEXT operation is implemented as a MPLS pop operation.

The INSERT operation is implemented as a MPLS push of a label stack.

The Node-SID value or Adj-SID value rightmost 20 bits MUST be used for label values. This implies SID values to be allocated according to the 20 bit space in MPLS labels.

5.3.3. IP Version 6

The text will be added in future revision.

6. IANA Considerations

TBD

7. Manageability Considerations

TBD

8. Security Considerations

TBD

9. Acknowledgements

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

10. References

10.1. Normative References

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

[RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.

[RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

[RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, October 2008.

[RFC5311] McPherson, D., Ginsberg, L., Previdi, S., and M. Shand, "Simplified Extension of Link State PDU (LSP) Space for IS-IS", RFC 5311, February 2009.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.

10.2. Informative References

[I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-01 (work in progress), December 2012.

[RFC5443] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, March 2009.

[RFC6138] Kini, S. and W. Lu, "LDP IGP Synchronization for Broadcast Networks", RFC 6138, February 2011.

Authors' Addresses

Stefano Previdi (editor)
Cisco Systems, Inc.
Via Del Serafico, 200
Rome 00142
Italy

Email: sprevidi@cisco.com

Clarence Filsfils (editor)
Cisco Systems, Inc.
Brussels
BE

Email: cfilsfil@cisco.com

Ahmed Bashandy
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: bashandy@cisco.com

Martin Horneffer
Deutsche Telekom
Hammer Str. 216-226
Muenster 48153
DE

Email: Martin.Horneffer@telekom.de

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange
FR

Email: stephane.litkowski@orange.com

Igor Milojevic
Telekom Srbija
Takovska 2
Belgrade
RS

Email: igormilojevic@telekom.rs

Rob Shakir
British Telecom
London
UK

Email: rob.shakir@bt.com

Saku Ytti
TDC Oy
Mechelininkatu 1a
TDC 00094
FI

Email: saku@ytti.fi

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
Antwerp 2018
BE

Email: wim.henderickx@alcatel-lucent.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: Jeff.Tantsura@ericsson.com