

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

C. Filsfils, Ed.
S. Previdi, Ed.
A. Bashandy
Cisco Systems, Inc.
B. Decraene
S. Litkowski
Orange
M. Horneffer
Deutsche Telekom
I. Milojevic
Telekom Srbija
R. Shakir
British Telecom
S. Ytti
TDC Oy
W. Henderickx
Alcatel-Lucent
J. Tantsura
Ericsson
E. Crabbe
Google, Inc.
July 3, 2014

Segment Routing Architecture
draft-filsfils-spring-segment-routing-04

Abstract

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through an ordered list of instructions, called segments. A segment can represent any instruction, topological or service-based. A segment can have a local semantic to an SR node or global within an SR domain. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node to the SR domain.

Segment Routing can be directly applied to the MPLS architecture with no change on the forwarding plane. A segment is encoded as an MPLS label. An ordered list of segments is encoded as a stack of labels. The segment to process is on the top of the stack. Upon completion of a segment, the related label is popped from the stack.

Segment Routing can be applied to the IPv6 architecture, with a new type of routing extension header. A segment is encoded as an IPv6 address. An ordered list of segments is encoded as an ordered list of IPv6 addresses in the routing extension header. The segment to

process is indicated by a pointer in the routing extension header. Upon completion of a segment, the pointer is incremented.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Companion Documents	4
2. Terminology	5
3. Link-State IGP Segments	7
3.1. IGP Segment, IGP SID	7
3.2. IGP-Prefix Segment, Prefix-SID	7
3.3. IGP-Node Segment, Node-SID	8
3.4. IGP-Anycast Segment, Anycast SID	9
3.5. IGP-Adjacency Segment, Adj-SID	9
3.5.1. Parallel Adjacencies	10
3.5.2. LAN Adjacency Segments	11
3.6. Binding Segment	11
3.6.1. Mapping Server	11
3.6.2. Tunnel Headend	11
3.6.3. Mirroring Context	12
3.7. Inter-Area Considerations	12
4. BGP Peering Segments	13
5. Multicast	14
6. IANA Considerations	14
7. Manageability Considerations	14
8. Security Considerations	14
9. Acknowledgements	14
10. References	14
10.1. Normative References	14
10.2. Informative References	14
Authors' Addresses	16

1. Introduction

With Segment Routing (SR), a node steers a packet through an ordered list of instructions, called segments. A segment can represent any instruction, topological or service-based. A segment can have a local semantic to an SR node or global within an SR domain. SR allows to enforce a flow through any path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

Segment Routing can be directly applied to the MPLS architecture (RFC 3031) with no change on the forwarding plane. A segment is encoded as an MPLS label. An ordered list of segments is encoded as a stack of labels. The active segment is on the top of the stack. A completed segment is popped off the stack. The addition of a segment is performed with a push.

In the Segment Routing MPLS instantiation, a segment could be of several types:

- o an IGP segment,
- o a BGP Peering segments,
- o an LDP LSP segment,
- o an RSVP-TE LSP segment,
- o a BGP LSP segment.

The first two (IGP and BGP Peering segments) types of segments defined in this document. The use of the last three types of segments is illustrated in [I-D.filsfils-spring-segment-routing-mpls].

Segment Routing can be applied to the IPv6 architecture (RFC2460), with a new type of routing extension header. A segment is encoded as an IPv6 address. An ordered list of segments is encoded as an ordered list of IPv6 addresses in the routing extension header. The active segment is indicated by a pointer in the routing extension header. Upon completion of a segment, the pointer is incremented. A segment can be inserted in the list and the pointer is updated accordingly.

Numerous use-cases illustrate the benefits of source routing either for FRR, OAM or Traffic Engineering reasons.

This document defines a set of instructions (called segments) that are required to fulfill the described use-cases. These segments can either be used in isolation (one single segment defines the source route of the packet) or in combination (these segments are part of an ordered list of segments that define the source route of the packet).

1.1. Companion Documents

This document defines the SR architecture, its routing model, the IGP-based segments, the BGP-based segments and the service segments.

Use cases are described in [I-D.filsfils-spring-segment-routing-use-cases], [I-D.ietf-spring-ipv6-use-cases], [I-D.ietf-spring-resiliency-use-cases], [I-D.geib-spring-oam-usecase] and [I-D.kumar-spring-sr-oam-requirement].

Segment Routing for MPLS dataplane is documented in [I-D.filsfils-spring-segment-routing-mpls].

Segment Routing for IPv6 dataplane is documented in [I-D.previdi-6man-segment-routing-header].

IGP protocol extensions for Segment Routing are described in [I-D.ietf-isis-segment-routing-extensions] and [I-D.ietf-ospf-segment-routing-extensions]

The FRR solution for SR is documented in [I-D.francois-segment-routing-ti-lfa].

The PCEP protocol extensions for Segment Routing are defined in [I-D.sivabalan-pce-segment-routing].

The interaction between SR/MPLS with other MPLS Signaling planes is documented in [I-D.filsfils-spring-segment-routing-ldp-interop].

2. Terminology

Segment: a segment identifies an instruction

SID: a Segment Identifier

Segment List: ordered list of SID's encoding the topological and service source route of the packet. It is a stack of labels in the MPLS architecture. It is an ordered list of IPv6 addresses in the IPv6 architecture.

Active segment: the segment that MUST be used by the receiving router to process the packet. It is identified by a pointer in the IPv6 architecture. It is the top label in the MPLS architecture.

PUSH: the insertion of a segment at the head of the Segment list.

NEXT: the active segment is completed, the next segment becomes active.

CONTINUE: the active segment is not completed and hence remains active. The CONTINUE instruction is implemented as the SWAP instruction in the MPLS dataplane.

SR Global Block (SRGB): local property of an SR node. In the MPLS architecture, SRGB is the set of local labels reserved for global segments. In the IPv6 architecture, it is the set of locally relevant IPv6 addresses.

Global Segment: the related instruction is supported by all the SR-capable nodes in the domain. In the MPLS architecture, a Global Segment has a globally-unique index. The related local label at a

given node N is found by adding the globally-unique index to the SRGB of node N. In the IPv6 architecture, a global segment is a globally-unique IPv6 address.

Local Segment: the related instruction is supported only by the node originating it. In the MPLS architecture, this is a local label outside the SRGB. In the IPv6 architecture, this is a link-local address.

IGP Segment: the generic name for a segment attached to a piece of information advertised by a link-state IGP, e.g. an IGP prefix or an IGP adjacency.

IGP-prefix Segment, Prefix-SID: an IGP-Prefix Segment is an IGP segment attached to an IGP prefix. An IGP-Prefix Segment is always global within the SR/IGP domain and identifies an instruction to forward the packet over the ECMP-aware shortest-path computed by the IGP to the related prefix. The Prefix-SID is the SID of the IGP-Prefix Segment.

IGP-Anycast: an IGP-Anycast Segment is an IGP-prefix segment which does not identify a specific router, but a set of routers. The terms "Anycast Segment" or "Anycast-SID" are often used as an abbreviation.

IGP-Adjacency: an IGP-Adjacency Segment is an IGP segment attached to an unidirectional adjacency or a set of unidirectional adjacencies. By default, an IGP-Adjacency Segment is local (unless explicitly advertised otherwise) to the node that advertises it.

IGP-Node: an IGP-Node Segment is an IGP-Prefix Segment which identifies a specific router (e.g. a loopback). The terms "Node Segment" or "Node-SID" are often used as an abbreviation.

SR Tunnel: a list of segments to be pushed on the packets directed on the tunnel. The list of segments can be specified explicitly or implicitly via a set of abstract constraints (latency, affinity, SRLG, ...). In the latter case, a constrained-based path computation is used to determine the list of segments associated with the tunnel. The computation can be local or delegated to a PCE server. An SR tunnel can be configured by the operator, provisioned via netconf or provisioned via PCEP. An SR tunnel can be used for traffic-engineering, OAM or FRR reasons.

Segment List Depth: the number of segments of an SR tunnel. The entity instantiating an SR Tunnel at a node N should be able to discover the depth insertion capability of the node N. The PCEP discovery capability is described in [I-D.sivabalan-pce-segment-routing].

3. Link-State IGP Segments

Within a link-state IGP domain, an SR-capable IGP node advertises segments for its attached prefixes and adjacencies. These segments are called IGP segments or IGP SIDs. They play a key role in Segment Routing and use-cases

([I-D.filsfils-spring-segment-routing-use-cases]) as they enable the expression of any topological path throughout the IGP domain. Such a topological path is either expressed as a single IGP segment or a list of multiple IGP segments.

3.1. IGP Segment, IGP SID

The terms "IGP Segment" and "IGP SID" are the generic names for a segment attached to a piece of information advertised by a link-state IGP, e.g. an IGP prefix or an IGP adjacency.

3.2. IGP-Prefix Segment, Prefix-SID

An IGP-Prefix Segment is an IGP segment attached to an IGP prefix. An IGP-Prefix Segment is always global within the SR/IGP domain and identifies the ECMP-aware shortest-path computed by the IGP to the related prefix. The Prefix-SID is the SID of the IGP-Prefix Segment.

A packet injected anywhere within the SR/IGP domain with an active Prefix-SID will be forwarded along the shortest-path to that prefix.

The IGP signaling extension for IGP-Prefix segment includes the P-Flag. A Node N advertising a Prefix-SID SID-R for its attached prefix R resets the P-Flag to allow its connected neighbors to perform the NEXT operation while processing SID-R. This behavior is equivalent to Penultimate Hop Popping in MPLS. When set, the neighbors of N must perform the CONTINUE operation while processing SID-R.

While SR allows to attach a local segment to an IGP prefix (using the L-Flag), we specifically assume that when the terms "IGP-Prefix Segment" and "Prefix-SID" are used, the segment is global (the SID is allocated from the SRGB). This is consistent with [I-D.filsfils-spring-segment-routing-use-cases] as all the described use-cases require global segments attached to IGP prefixes.

A single Prefix-SID is allocated to an IGP Prefix in a topology.

In the context of multiple topologies, multiple Prefix-SID's MAY be allocated to the same IGP Prefix (e.g.: using the "algorithm" field in the IGP advertisement as described in [I-D.ietf-isis-segment-routing-extensions] and

[I-D.ietf-ospf-segment-routing-extensions])). However, each prefix-SID MUST be associated with only one topology. In other words: a prefix, within a topology, MUST have only a single Prefix-SID.

A Prefix-SID is allocated from the SRGB according to a process similar to IP address allocation. Typically the Prefix-SID is allocated by policy by the operator (or NMS) and the SID very rarely changes.

The allocation process MUST NOT allocate the same Prefix-SID to different IP prefixes.

If a node learns a Prefix-SID having a value that falls outside the locally configured SRGB range, then the node MUST NOT use the Prefix-SID and SHOULD issue an error log warning for misconfiguration.

The required IGP protocol extensions are defined in [I-D.ietf-isis-segment-routing-extensions] and [I-D.ietf-ospf-segment-routing-extensions].

A node N attaching a Prefix-SID SID-R to its attached prefix R MUST maintain the following FIB entry:

```
Incoming Active Segment: SID-R
Ingress Operation: NEXT
Egress interface: NULL
```

A remote node M MUST maintain the following FIB entry for any learned Prefix-SID SID-R attached to IP prefix R:

```
Incoming Active Segment: SID-R
Ingress Operation:
  If the next-hop of R is the originator of R
  and instructed to remove the active segment: NEXT
  Else: CONTINUE
Egress interface: the interface towards the next-hop along
                  the shortest-path to prefix R.
```

3.3. IGP-Node Segment, Node-SID

An IGP-Node Segment is a an IGP-Prefix Segment which identifies a specific router (e.g. a loopback). The N flag is set. The terms "Node Segment" or "Node-SID" are often used as an abbreviation.

A "Node Segment" or "Node-SID" is fundamental to SR. From anywhere in the network, it enforces the ECMP-aware shortest- path forwarding of the packet towards the related node as explained in [I-D.filsfils-spring-segment-routing-use-cases].

An IGP-Node-SID MUST NOT be associated with a prefix that is owned or advertised by more than one router within the same routing domain.

3.4. IGP-Anycast Segment, Anycast SID

An IGP-Anycast Segment is an IGP-prefix segment which does not identify a specific router, but a set of routers. The terms "Anycast Segment" or "Anycast-SID" are often used as an abbreviation.

An "Anycast Segment" or "Anycast SID" enforces the ECMP-aware shortest-path forwarding towards the closest node of the anycast set. This is useful to express macro-engineering policies or protection mechanisms as described in [I-D.filsfils-spring-segment-routing-use-cases].

The Anycast SID MUST be advertised with the N-flag unset.

3.5. IGP-Adjacency Segment, Adj-SID

An IGP-Adjacency Segment is an IGP segment attached to a unidirectional adjacency or a set of unidirectional adjacencies. By default, an IGP-Adjacency Segment is local to the node which advertises it. However, an Adjacency Segment can be global if advertised by the IGP as such. The SID of the IGP-Adjacency Segment is called the Adj-SID.

The adjacency is formed by the local node (i.e., the node advertising the adjacency in the IGP) and the remote node (i.e., the other end of the adjacency). The local node MUST be an IGP node. The remote node MAY be an adjacent IGP neighbor) or a non-adjacent neighbor (e.g.: a Forwarding Adjacency, [RFC4206]).

A packet injected anywhere within the SR domain with a segment list {SN, SNL}, where SN is the Node-SID of node N and SNL is an Adj-SID attached by node N to its adjacency over link L, will be forwarded along the shortest-path to N and then be switched by N, without any IP shortest-path consideration, towards link L. If the Adj-SID identifies a set of adjacencies, then the node N load-balances the traffic among the various members of the set.

Similarly, when using a global Adj-SID, a packet injected anywhere within the SR domain with a segment list {SNL}, where SNL is a global Adj-SID attached by node N to its adjacency over link L, will be forwarded along the shortest-path to N and then be switched by N, without any IP shortest-path consideration, towards link L. If the Adj-SID identifies a set of adjacencies, then the node N load-balances the traffic among the various members of the set. The use of global Adj-SID allows to reduce the size of the segment list when

expressing a path at the cost of additional state (i.e.: the global Adj-SID will be inserted by all routers within the area in their forwarding table).

An "IGP Adjacency Segment" or "Adj-SID" enforces the switching of the packet from a node towards a defined interface or set of interfaces. This is key to theoretically prove that any path can be expressed as a list of segments as explained in [I-D.filsfils-spring-segment-routing-use-cases].

The encodings of the Adj-SID include the B-flag. When set, the Adj-SID benefits from a local protection.

The encodings of the Adj-SID include the L-flag. When set, the Adj-SID has local significance. By default the L-flag is set.

A node SHOULD allocate one Adj-SIDs for each of its adjacencies.

A node MAY allocate multiple Adj-SIDs to the same adjacency. An example is where the adjacency is established over a bundle interface. Each bundle member MAY have its own Adj-SID.

A node MAY allocate the same Adj-SID to multiple adjacencies.

Adjacency suppression MUST NOT be performed by the IGP.

A node MUST install a FIB entry for any Adj-SID of value V attached to data-link L:

```
Incoming Active Segment: V
Operation: NEXT
Egress Interface: L
```

The Adj-SID implies, from the router advertising it, the forwarding of the packet through the adjacency identified by the Adj-SID, regardless its IGP/SPF cost. In other words, the use of Adjacency Segments overrides the routing decision made by SPF algorithm.

3.5.1. Parallel Adjacencies

Adj-SIDs can be used in order to represent a set of parallel interfaces between two adjacent routers.

A node MUST install a FIB entry for any locally originated Adjacency Segment (Adj-SID) of value W attached to a set of link B with:

Incoming Active Segment: W
Ingress Operation: NEXT
Egress interface: loadbalance between any data-link within set B

3.5.2. LAN Adjacency Segments

In LAN subnetworks, link-state protocols define the concept of Designated Router (DR, in OSPF) or Designated Intermediate System (DIS, in IS-IS) that conduct flooding in broadcast subnetworks and that describe the LAN topology in a special routing update (OSPF Type2 LSA or IS-IS Pseudonode LSP).

The difficulty with LANs is that each router only advertises its connectivity to the DR/DIS and not to each other individual nodes in the LAN. Therefore, additional protocol mechanisms (IS-IS and OSPF) are necessary in order for each router in the LAN to advertise an Adj-SID associated to each neighbor in the LAN. These extensions are defined in [I-D.ietf-isis-segment-routing-extensions] and [I-D.ietf-ospf-segment-routing-extensions].

3.6. Binding Segment

3.6.1. Mapping Server

A Remote-Binding SID S advertised by the mapping server M for remote prefix R attached to non-SR-capable node N signals the same information as if N had advertised S as a Prefix-SID. Further details are described in the SR/LDP interworking procedures ([I-D.filsfils-spring-segment-routing-ldp-interop]).

The segment allocation and SRDB Maintenance rules are the same as those defined for Prefix-SID.

3.6.2. Tunnel Headend

The segment allocation and SRDB Maintenance rules are the same as those defined for Adj-SID. A tunnel attached to a head-end H acts as an adjacency attached to H.

Note: an alternative would consist in representing tunnels as forwarding-adjacencies ([RFC4206]). The Remote-Binding SID is preferred as it allows to advertise the presence of a tunnel without influencing the LSDB and the SPF computation.

3.6.3. Mirroring Context

TBD.

3.7. Inter-Area Considerations

In the following example diagram we assume an IGP deployed using areas and where SR has been deployed.

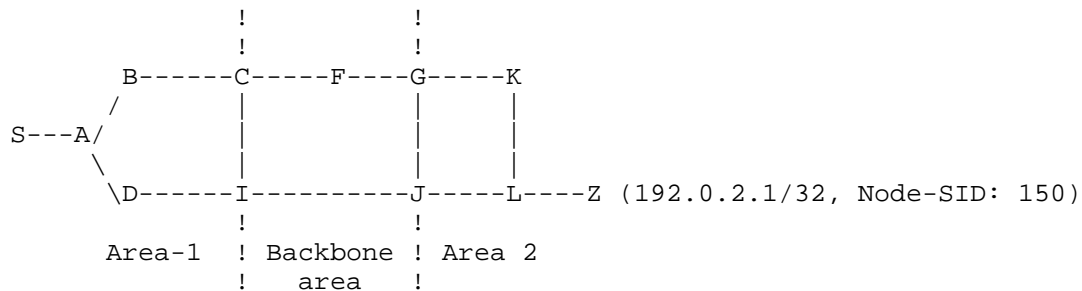


Figure 1: Inter-Area Topology Example

In area 2, node Z allocates Node-SID 150 to his local prefix 192.0.2.1/32. ABRs G and J will propagate the prefix into the backbone area by creating a new instance of the prefix according to normal inter-area/level IGP propagation rules.

Nodes C and I will apply the same behavior when leaking prefixes from the backbone area down to area 1. Therefore, node S will see prefix 192.0.2.1/32 with Prefix-SID 150 and advertised by nodes C and I.

It therefore results that a Prefix-SID remains attached to its related IGP Prefix through the inter-area process.

When node S sends traffic to 192.0.2.1/32, it pushes Node-SID(150) as active segment and forwards it to A.

When packet arrives at ABR I (or C), the ABR forwards the packet according to the active segment (Node-SID(150)). Forwarding continues across area borders, using the same Node-SID(150), until the packet reaches its destination.

When an ABR propagates a prefix from one area to another it MUST set the R-Flag.

4. BGP Peering Segments

In the context of BGP Egress Peer Engineering (EPE), as described in [draft-filsfils-spring-segment-routing-central-epe], an EPE enabled Egress PE node MAY advertise segments corresponding to its attached peers. These segments are called BGP peering segments or BGP Peering SIDs. They enable the expression of source-routed inter-domain paths.

An ingress border router of an AS may compose a list of segments to steer a flow along a selected path within the AS, towards a selected egress border router C of the AS and through a specific peer. At minimum, a BGP Peering Engineering policy applied at an ingress PE involves two segments: the Node SID of the chosen egress PE and then the BGP Peering Segment for the chosen egress PE peer or peering interface.

Hereafter, we will define three types of BGP peering segments/SID's: PeerNodeSID, PeerAdjSID and PeerSetSID.

- o PeerNode SID. A BGP PeerNode segment/SID is a local segment. At the BGP node advertising it, its semantics is:
 - * SR header operation: NEXT.
 - * Next-Hop: the connected peering node to which the segment is related.
- o PeerAdj SID: A BGP PeerAdj segment/SID is a local segment. At the BGP node advertising it, its semantics is:
 - * SR header operation: NEXT.
 - * Next-Hop: the peer connected through the interface to which the segment is related.
- o PeerSet SID. A BGP PeerSet segment/SID is a local segment. At the BGP node advertising it, its semantics is:
 - * SR header operation: NEXT.
 - * Next-Hop: loadbalance across any connected interface to any peer in the related group.

A peer set could be all the connected peers from the same AS or a subset of these. A group could also span across AS. The group definition is a policy set by the operator.

The BGP extensions necessary in order to signal these BGP peering segments will be defined in a separate document.

5. Multicast

Segment Routing is defined for unicast. The application of the source-route concept to Multicast is not in the scope of this document.

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, Les Ginsberg, Ruediger Geib and Hannes Gredler 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.

10.2. Informative References

- [I-D.filsfils-spring-segment-routing-ldp-interop] Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing interoperability with LDP", draft-filsfils-spring-segment-routing-ldp-interop-01 (work in progress), April 2014.

[I-D.filsfils-spring-segment-routing-mpls]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-filsfils-spring-segment-routing-mpls-02 (work in progress), June 2014.

[I-D.filsfils-spring-segment-routing-use-cases]

Filsfils, C., Francois, P., Previdi, S., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., Kini, S., and E. Crabbe, "Segment Routing Use Cases", draft-filsfils-spring-segment-routing-use-cases-00 (work in progress), March 2014.

[I-D.francois-segment-routing-ti-lfa]

Francois, P., Filsfils, C., Bashandy, A., and B. Decraene, "Topology Independent Fast Reroute using Segment Routing", draft-francois-segment-routing-ti-lfa-00 (work in progress), November 2013.

[I-D.geib-spring-oam-usecase]

Geib, R. and C. Filsfils, "Use case for a scalable and topology aware MPLS data plane monitoring system", draft-geib-spring-oam-usecase-01 (work in progress), February 2014.

[I-D.ietf-isis-segment-routing-extensions]

Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-02 (work in progress), June 2014.

[I-D.ietf-ospf-segment-routing-extensions]

Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-00 (work in progress), June 2014.

[I-D.ietf-spring-ipv6-use-cases]

Brzozowski, J., Leddy, J., Leung, I., Previdi, S., Townsley, W., Martin, C., Filsfils, C., and R. Maglione, "IPv6 SPRING Use Cases", draft-ietf-spring-ipv6-use-cases-00 (work in progress), May 2014.

[I-D.ietf-spring-resiliency-use-cases]

Francois, P., Filsfils, C., Decraene, B., and R. Shakir,
"Use-cases for Resiliency in SPRING", draft-ietf-spring-
resiliency-use-cases-00 (work in progress), May 2014.

[I-D.kumar-spring-sr-oam-requirement]

Kumar, N., Pignataro, C., Akiya, N., Geib, R., and G.
Mirsky, "OAM Requirements for Segment Routing Network",
draft-kumar-spring-sr-oam-requirement-00 (work in
progress), February 2014.

[I-D.previdi-6man-segment-routing-header]

Previdi, S., Filsfils, C., Field, B., and I. Leung, "IPv6
Segment Routing Header (SRH)", draft-previdi-6man-segment-
routing-header-01 (work in progress), June 2014.

[I-D.sivabalan-pce-segment-routing]

Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and
R. Raszuk, "PCEP Extensions for Segment Routing", draft-
sivabalan-pce-segment-routing-02 (work in progress),
October 2013.

[draft-filsfils-spring-segment-routing-central-epe]

Filsfils, C. and S. Previdi, "Segment Routing Centralized
Egress Peer Engineering", May 2013.

Authors' Addresses

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

Email: cfilsfil@cisco.com

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

Email: sprevidi@cisco.com

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

Email: bashandy@cisco.com

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange
FR

Email: stephane.litkowski@orange.com

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

Email: Martin.Horneffer@telekom.de

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

Edward Crabbe
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: edc@google.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 5, 2015

C. Filsfils, Ed.
S. Previdi, Ed.
K. Patel
Cisco Systems, Inc.
E. Aries
S. Shaw
Facebook
D. Ginsburg
D. Afanasiev
Yandex
July 4, 2014

Segment Routing Centralized Egress Peer Engineering
draft-filsfils-spring-segment-routing-central-epe-02

Abstract

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document illustrates the application of Segment Routing to solve the Egress Peer Engineering (EPE) requirement. The SR-based EPE solution allows a centralized (SDN) controller to program any egress peer policy at ingress border routers or at hosts within the domain. This document is on the informational track.

Requirements Language

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

Status of This Memo

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

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

working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Segment Routing Documents	4
1.2. Problem Statement	4
2. BGP Peering Segments	6
3. Distribution of External Topology and TE Information using BGP-LS	6
3.1. EPE Route advertising the Peer D and its PeerNode SID	7
3.2. EPE Route advertising the Peer E and its PeerNode SID	7
3.3. EPE Route advertising the Peer F and its PeerNode SID	8
3.4. EPE Route advertising a first PeerAdj to Peer F	8
3.5. EPE Route advertising a second PeerAdj to Peer F	9
3.6. FRR	9
4. EPE Controller	10
4.1. Valid Paths From Peers	11
4.2. Intra-Domain Topology	11
4.3. External Topology	11
4.4. SLA characteristics of each peer	12
4.5. Traffic Matrix	12
4.6. Business Policies	12
4.7. EPE Policy	12
5. Programming an input policy	13

5.1. At a Host	13
5.2. At a router - SR Traffic Engineering tunnel	13
5.3. At a Router - BGP3107 policy route	14
5.4. At a Router - VPN policy route	14
5.5. At a Router - Flowspec route	14
6. IPv6	15
7. Benefits	15
8. IANA Considerations	16
9. Manageability Considerations	16
10. Security Considerations	16
11. Acknowledgements	16
12. References	16
12.1. Normative References	16
12.2. Informative References	16
Authors' Addresses	18

1. Introduction

The document is structured as follows:

- o Section 1 reminds the EPE problem statement and provides the key references.
- o Section 2 defines the different BGP Peering Segments and the semantic associated to them.
- o Section 3 describes the automated allocation of BGP Peering SID's by the EPE-enabled egress border router and the automated signaling of the external peering topology and the related BGP Peering SID's to the collector
[[I-D.previdi-idr-bgpls-segment-routing-epe].
- o Section 4 overviews the components of a centralized EPE controller. The definition of the EPE controller is outside the scope of this document.
- o Section 5 overviews the methods that could be used by the centralized EPE controller to implement an EPE policy at an ingress border router or at a source host within the domain. The exhaustive definition of all the means to program an EPE input policy is outside the scope of this document.

For editorial reason, the solution is described for IPv4. A later section describes how the same solution is applicable to IPv6.

1.1. Segment Routing Documents

The main references for this document are:

- o SR Problem Statement: [I-D.ietf-spring-problem-statement].
- o SR Architecture: [I-D.filsfils-spring-segment-routing].
- o Distribution of External Topology and TE Information using BGP: [I-D.previdi-idr-bgpls-segment-routing-epe].

The SR instantiation in the MPLS dataplane is described in [I-D.filsfils-spring-segment-routing-mpls].

The SR IGP protocol extensions are defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.psenak-ospf-segment-routing-ospfv3-extension].

The Segment Routing PCE protocol extensions are defined in [I-D.sivabalan-pce-segment-routing].

1.2. Problem Statement

The EPE problem statement is defined in [I-D.ietf-spring-problem-statement].

A centralized controller should be able to instruct an ingress PE or a content source within the domain to use a specific egress PE and a specific external interface to reach a particular destination.

We call this solution "EPE" for "Egress Peer Engineering". The centralized controller is called the "EPE Controller". The egress border router where the EPE traffic-steering functionality is implemented is called an EPE-enabled border router. The input policy programmed at an ingress border router or at a source host is called an EPE policy.

The requirements that have motivated the solution described in this document are listed here below:

- o The solution MUST apply to the Internet use-case where the Internet routes are assumed to use IPv4 unlabeled or IPv6 unlabeled. It is not required to place the internet routes in a VRF and allocate labels on a per route, or on a per-path basis.

- o The solution MUST NOT make any assumption on the currently deployed iBGP schemes (RRs, confederations or iBGP full meshes) and MUST be able to support all of them.
- o The solution SHOULD minimize the need for new BGP capabilities at the ingress PE's.
- o The solution MUST accommodate an ingress EPE policy at an ingress PE or directly at a source host within the domain.
- o The solution MUST support automated FRR and fast convergence.

The following reference diagram is used throughout this document.

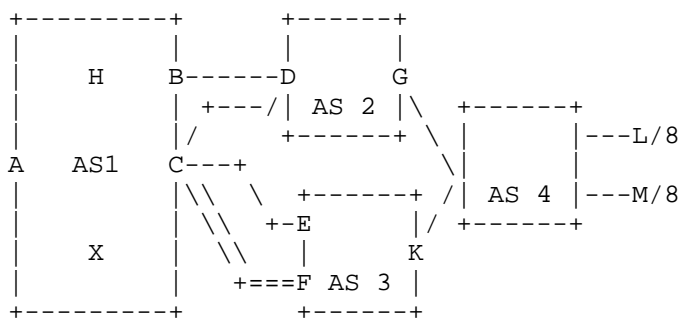


Figure 1: Reference Diagram

IPv4 addressing:

- o C's interface to D: 1.0.1.1/24, D's interface: 1.0.1.2/24
- o C's interface to E: 1.0.2.1/24, E's interface: 1.0.2.2/24
- o C's upper interface to F: 1.0.3.1/24, F's interface: 1.0.3.2/24
- o C's lower interface to F: 1.0.4.1/24, F's interface: 1.0.4.2/24
- o Loopback of F used for eBGP multi-hop peering to C: 1.0.5.2/32
- o C's loopback is 3.3.3.3/32 with SID 64

C's BGP peering:

- o Single-hop eBGP peering with neighbor 1.0.1.2 (D)
- o Single-hop eBGP peering with neighbor 1.0.2.2 (E)

- o Multi-hop eBGP peering with F on ip address 1.0.5.2 (F)

C's resolution of the multi-hop eBGP session to F:

- o Static route 1.0.5.2/32 via 1.0.3.2
- o Static route 1.0.5.2/32 via 1.0.4.2

C is configured with local policy that defines a BGP PeerSet as the set of peers (1.0.2.2 and 1.0.5.2)

X is the EPE controller within AS1 domain.

H is a content source within AS1 domain.

2. BGP Peering Segments

AS defined in [I-D.filsfils-spring-segment-routing], Segments are defined by a Egress Peer Engineering (EPE) capable node and corresponding to its attached peers. These segments are called BGP peering segments or BGP Peering SIDs. They enable the expression of source-routed inter-domain paths.

An ingress border router of an AS may compose a list of segments to steer a flow along a selected path within the AS, towards a selected egress border router C of the AS and through a specific peer. At minimum, a BGP Peering Engineering policy applied at an ingress PE involves two segments: the Node SID of the chosen egress PE and then the BGP Peering Segment for the chosen egress PE peer or peering interface.

[I-D.filsfils-spring-segment-routing] defines three types of BGP peering segments/SID's: PeerNodeSID, PeerAdjSID and PeerSetSID.

The BGP extensions to signal these BGP peering segments are outlined in the following section.

3. Distribution of External Topology and TE Information using BGP-LS

In ships-in-the-night mode with respect to the pre-existing iBGP design, a BGPLS session is established between the EPE-enabled border router and the EPE controller.

As a result of its local configuration and according to the behavior described in [I-D.previdi-idr-bgpls-segment-routing-epe], node C allocates the following BGP Peering Segments ([I-D.filsfils-spring-segment-routing]):

- o A PeerNode segment for each of its defined peer (D, E and F).
- o A PeerAdj segment for each recursing interface to a multi-hop peer (e.g.: the upper and lower interfaces from C to F in figure 1).
- o A PeerSet segment to the set of peers (E and F).

C programs its forwarding table accordingly:

Incoming Label	Operation	Outgoing Interface

1012	POP	link to D
1022	POP	link to E
1032	POP	upper link to F
1042	POP	lower link to F
1052	POP	loadbalance on any link to F
1060	POP	loadbalance on any link to E or to F

C signals the related BGP-LS NLRI's to the EPE controller. Each such BGP-LS route is described in the following sub-sections according to the encoding details defined in draft-previdi-idr-bgpl-segment-routing-epe-00.

3.1. EPE Route advertising the Peer D and its PeerNode SID

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS2
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
1.0.1.1, 1.0.1.2

Attributes:

- o Adj-SID: 1012

3.2. EPE Route advertising the Peer E and its PeerNode SID

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3

- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
1.0.2.1, 1.0.2.2

Attributes:

- o Adj-SID: 1022
- o PeerSetSID: 1060
- o Link Attributes: see section 3.3.2 of
[I-D.ietf-idr-ls-distribution]

3.3. EPE Route advertising the Peer F and its PeerNode SID

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
3.3.3.3, 1.0.5.2

Attributes:

- o Adj-SID: 1052
- o PeerSetSID: 1060

3.4. EPE Route advertising a first PeerAdj to Peer F

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
1.0.3.1 , 1.0.3.2

Attributes:

- o Adj-SID: 1032
- o LinkAttributes: see section 3.3.2 of
[I-D.ietf-idr-ls-distribution]

3.5. EPE Route advertising a second PeerAdj to Peer F

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
1.0.4.1 , 1.0.4.2

Attributes:

- o Adj-SID: 1042
- o LinkAttributes: see section 3.3.2 of
[I-D.ietf-idr-ls-distribution]

3.6. FRR

An EPE-enabled border router should allocate a FRR backup entry on a per BGP Peering SID basis:

- o PeerNode SID
 1. If multi-hop, backup via the remaining PeerADJ SID's to the same peer.
 2. Else backup via local PeerNode SID to the same AS.
 3. Else pop the PeerNode SID and IP lookup (with potential BGP PIC fall-back).
- o PeerAdj SID
 1. If to a multi-hop peer, backup via the remaining PeerADJ SID's to the same peer.
 2. Else backup via PeerNode SID to the same AS.
 3. Else pop the PeerNode SID and IP lookup (with potential BGP PIC fall-back).
- o PeerSet SID
 1. Backup via remaining PeerNode SID in the same PeerSet.

2. Else pop the PeerSet SID and IP lookup (with potential BGP PIC fall-back).

We illustrate the different types of possible backups using the reference diagram and considering the Peering SID's allocated by C.

PeerNode SID 1052, allocated by C for peer F:

- o Upon the failure of the upper connected link CF, C can reroute all the traffic onto the lower CF link to the same peer (F).

PeerNode SID 1022, allocated by C for peer E:

- o Upon the failure of the connected link CE, C can reroute all the traffic onto the link to PeerNode SID 1052 (F).

PeerNode SID 1012, allocated by C for peer D:

- o Upon the failure of the connected link CD, C can pop the PeerNode SID and lookup the IP destination address in its FIB and route accordingly.

PeerSet SID 1060, allocated by C for the set of peers E and F:

- o Upon the failure of a connected link in the group, the traffic to PeerSet SID 1060 is rerouted on any other member of the group.

For specific business reasons, the operator might not want the default FRR behavior applied to a PeerNode SID or any of its depending PeerADJ SID.

The operator should be able to associate a specific backup PeerNode SID for a PeerNode SID: e.g. 1022 (E) must be backed up by 1012 (D) which over-rules the default behavior which would have preferred F as a backup for E.

4. EPE Controller

In this section, we provide a non-exhaustive set of inputs that an EPE controller would likely collect such as to perform the EPE policy decision.

The exhaustive definition is outside the scope of this document.

4.1. Valid Paths From Peers

The EPE controller should collect all the paths advertised by all the engineered peers.

This could be realized by setting an iBGP session with the EPE-enabled border router, with "add-path all" and original next-hop preserved.

In this case, C would advertise the following Internet routes to the EPE controller:

- o NLRI <L/8>, nhop 1.0.1.2, AS Path {AS 2, 4}
 - * X (i.e.: the EPE controller) knows that C receives a path to L/8 via neighbor 1.0.1.2 of AS2.
- o NLRI <L/8>, nhop 1.0.2.2, AS Path {AS 3, 4}
 - * X knows that C receives a path to L/8 via neighbor 1.0.2.2 of AS2.
- o NLRI <L/8>, nhop 1.0.5.2, AS Path {AS 3, 4}
 - * X knows that C has an eBGP path to L/8 via AS3 via neighbor 1.0.5.2

An alternative option consists in Adj-RIB-In BMP from EPE-enabled border router to the EPE collector.

4.2. Intra-Domain Topology

The EPE controller should collect the internal topology and the related IGP SID's.

This could be realized by collecting the IGP LSDB of each area or running a BGP-LS session with a node in each IGP area.

4.3. External Topology

Thanks to the collected BGP-LS routes described in the section 2 (BGPLS advertisements), the EPE controller is able to maintain an accurate description of the egress topology of node C. Furthermore, the EPE controller is able to associate BGP Peering SID's to the various components of the external topology.

4.4. SLA characteristics of each peer

The EPE controller might collect SLA characteristics across peers. This requires an EPE solution as the SL A probes need to be steered via non-best-path peers.

Uni-directional SLA monitoring of the desired path is likely required. This might be possible when the application is controlled at the source and the receiver side. Uni-directional monitoring dissociates the SLA characteristic of the return path (which cannot usually be controlled) from the forward path (the one of interest for pushing content from a source to a consumer and the one which can be controlled).

Alternatively, Extended Metrics, as defined in [I-D.ietf-isis-te-metric-extensions] could also be advertised using new bgpls attributes.

4.5. Traffic Matrix

The EPE controller might collect the traffic matrix to its peers or the final destinations. IPFIX is a likely option.

An alternative option consists in collecting the link utilization statistics of each of the internal and external links, also available in current definition of [I-D.ietf-idr-ls-distribution].

4.6. Business Policies

The EPE controller should collect business policies.

4.7. EPE Policy

On the basis of all these inputs (and likely other), the EPE Controller decides to steer some demands away from their best BGP path.

The EPE policy is likely expressed as a two-entry segment list where the first element is the IGP prefix SID of the selected egress border router and the second element is a BGP Peering SID at the selected egress border router.

A few examples are provided hereafter:

- o Prefer egress PE C and peer AS AS2: {64, 1012}.
- o Prefer egress PE C and peer AS AS3 via ebgp peer 1.0.2.2: {64, 1022}.

- o Prefer egress PE C and peer AS AS3 via ebgp peer 1.0.5.2: {64, 1052}.
- o Prefer egress PE C and peer AS AS3 via interface 1.0.4.2 of multi-hop ebgp peer 1.0.5.2: {64, 1042}.
- o Prefer egress PE C and any interface to any peer in the group 1060: {64, 1060}.

Note that the first SID could be replaced by a list of segments. This is useful when an explicit path within the domain is required for traffic-engineering purpose. For example, if the Prefix SID of node B is 60 and the EPE controller would like to steer the traffic from A to C via B then through the external link to peer D then the segment list would be {60, 64, 1012}.

5. Programming an input policy

The detailed/exhaustive description of all the means to implement an EPE policy are outside the scope of this document. A few examples are provided in this section.

5.1. At a Host

A static IP/MPLS route can be programmed at the host H. The static route would define a destination prefix, a next-hop and a label stack to push. The global property of the IGP Prefix SID is particularly convenient: the same policy could be programmed across hosts connected to different routers.

5.2. At a router - SR Traffic Engineering tunnel

The EPE controller can configure the ingress border router with an SR traffic engineering tunnel T1 and a steering-policy S1 which causes a certain class of traffic to be mapped on the tunnel T1.

The tunnel T1 would be configured to push the require segment list.

The tunnel and the steering policy could be configured via PCEP according to [I-D.sivabalan-pce-segment-routing] and [I-D.ietf-pce-pce-initiated-lsp] or via Netconf ([RFC6241]).

Example: at A

```
Tunnel T1: push {64, 1042}
IP route L/8 set nhop T1
```

5.3. At a Router - BGP3107 policy route

The EPE Controller could build a BGP3107 ([RFC3107]) route (from scratch) and send it to the ingress router:

- o NLRI: the destination prefix to engineer: e.g. L/8.
- o Next-Hop: the selected egress border router: C.
- o Label: the selected egress peer: 1042.
- o AS path: reflecting the valid AS path of the selected.
- o Some BGP policy to ensure it be selected as best by the ingress router.

This BGP3107 policy route "overwrites" an equivalent or less-specific "best path". As the best-path is changed, this EPE input policy option influences the path propagated to the upstream peer/customers.

5.4. At a Router - VPN policy route

The EPE Controller could build a VPNv4 route (from scratch) and send it to the ingress router:

- o NLRI: the destination prefix to engineer: e.g. L/8.
- o Next-Hop: the selected egress border router: C.
- o Label: the selected egress peer: 1042.
- o Route-Target: selecting the appropriate VRF at the ingress router.
- o AS path: reflecting the valid AS path of the selected.
- o Some BGP policy to ensure it be selected as best by the ingress router in the related VRF.

The related VRF must be pre-configured. A VRF fall-back into main FIB might be beneficial to avoid replicating all the "normal" internet paths in each VRF.

5.5. At a Router - Flowspec route

EPE Controller builds a FlowSpec route and sends it to the ingress router to engineer:

- o Dissemination of Flow Specification Rules ([RFC5575]).

- o Destination/Source IP Addresses, IP Protocol, Destination/Source port (+1 component).
- o ICMP Type/Code, TCP Flags, Packet length, DSCP, Fragment.

6. IPv6

The described solution is applicable to IPv6, either with MPLS-based or IPv6-Native segments. In both cases, the same three steps of the solution are applicable:

- o BGP-LS-based signaling of the external topology and BGP Peering Segments to the EPE controller.
- o Collection of various inputs by the EPE controller to come up with a policy decision.
- o Programming at an ingress router or source host of the desired EPE policy which consists in a list of segments to push on a defined traffic class.

7. Benefits

The EPE solutions described in this document has the following benefits:

- o No assumption on the iBGP design with AS1.
- o Next-Hop-Self on the internet routes propagated to the ingress border routers is possible. This is a common design rule to minimize the number of IGP routes and to avoid importing external churn into the internal domain.
- o Consistent support for traffic-engineering within the domain and at the external edge of the domain.
- o Support host and ingress border router EPE policy programming.
- o EPE functionality is only required on the EPE-enabled egress border router and the EPE controller: an ingress policy can be programmed at the ingress border router without any new functionality.
- o Ability to deploy the same input policy across hosts connected to different routers (global property of the IGP prefix SID).

8. IANA Considerations

TBD

9. Manageability Considerations

TBD

10. Security Considerations

TBD

11. Acknowledgements

TBD

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.
- [RFC6241] Enns, R., Bjorklund, M., Schoenwaelder, J., and A. Bierman, "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011.

12.2. Informative References

- [I-D.filsfils-spring-segment-routing] Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-spring-segment-routing-03 (work in progress), June 2014.

- [I-D.filsfils-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-filsfils-spring-segment-routing-mpls-02 (work in progress), June 2014.
- [I-D.ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-05 (work in progress), May 2014.
- [I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-02 (work in progress), June 2014.
- [I-D.ietf-isis-te-metric-extensions]
Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., Filsfils, C., and W. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-ietf-isis-te-metric-extensions-03 (work in progress), April 2014.
- [I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-00 (work in progress), June 2014.
- [I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-01 (work in progress), June 2014.
- [I-D.ietf-spring-problem-statement]
Previdi, S., Filsfils, C., Decraene, B., Litkowski, S., Horneffer, M., Geib, R., Shakir, R., and R. Raszuk, "SPRING Problem Statement and Requirements", draft-ietf-spring-problem-statement-01 (work in progress), June 2014.

- [I-D.previdi-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., and K. Patel, "Segment Routing Egress Peer Engineering BGPLS Extensions", draft-previdi-idr-bgpls-segment-routing-epe-00 (work in progress), May 2014.
- [I-D.psenak-ospf-segment-routing-ospfv3-extension]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPFv3 Extensions for Segment Routing", draft-psenak-ospf-segment-routing-ospfv3-extension-02 (work in progress), July 2014.
- [I-D.sivabalan-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and R. Raszuk, "PCEP Extensions for Segment Routing", draft-sivabalan-pce-segment-routing-02 (work in progress), October 2013.

Authors' Addresses

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

Email: cfilsfil@cisco.com

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

Email: sprevidi@cisco.com

Keyur Patel
Cisco Systems, Inc.
US

Email: keyupate@cisco.com

Ebben Aries
Facebook
US

Email: exa@fb.com

Steve Shaw
Facebook
US

Email: shaw@fb.com

Daniel Ginsburg
Yandex
RU

Email: dbg@yandex-team.ru

Dmitry Afanasiev
Yandex
RU

Email: fl0w@yandex-team.ru

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 8, 2014

C. Filsfils, Ed.
S. Previdi, Ed.
A. Bashandy
Cisco Systems, Inc.
B. Decraene
S. Litkowski
Orange
M. Horneffer
Deutsche Telekom
I. Milojevic
Telekom Srbija
R. Shakir
British Telecom
S. Ytti
TDC Oy
W. Henderickx
Alcatel-Lucent
J. Tantsura
Ericsson
E. Crabbe
Google, Inc.
June 6, 2014

Segment Routing with MPLS data plane
draft-filsfils-spring-segment-routing-mpls-02

Abstract

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node to the SR domain.

Segment Routing can be directly applied to the MPLS architecture with no change in the forwarding plane. This drafts describes how Segment Routing operates on top of the MPLS data plane.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on December 8, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Illustration	3
3. MPLS Instantiation of Segment Routing	4
4. IGP Segments Examples	5
4.1. Example 1	6
4.2. Example 2	7
4.3. Example 3	7
4.4. Example 4	7
4.5. Example 5	8
5. Other Examples of MPLS Segments	8
5.1. LDP LSP segment combined with IGP segments	8
5.2. RSVP-TE LSP segment combined with IGP segments	9
6. Segment List History	10
7. IANA Considerations	10

8. Manageability Considerations	10
9. Security Considerations	10
10. Acknowledgements	11
11. References	11
11.1. Normative References	11
11.2. Informative References	11
Authors' Addresses	12

1. Introduction

The Segment Routing architecture [I-D.filsfils-rtgwg-segment-routing] can be directly applied to the MPLS architecture with no change in the MPLS forwarding plane. This drafts describes how Segment Routing operates on top of the MPLS data plane.

The Segment Routing use cases are described in in [I-D.filsfils-rtgwg-segment-routing-use-cases].

Link State protocol extensions for Segment Routing are described in [I-D.previdi-isis-segment-routing-extensions], [I-D.psenak-ospf-segment-routing-extensions] and [I-D.psenak-ospf-segment-routing-ospfv3-extension].

2. Illustration

Segment Routing, applied to the MPLS data plane, offers the ability to tunnel services (VPN, VPLS, VPWS) from an ingress PE to an egress PE, without any other protocol than ISIS or OSPF ([I-D.previdi-isis-segment-routing-extensions] and [I-D.psenak-ospf-segment-routing-extensions]). LDP and RSVP-TE signaling protocols are not required.

Note that [draft-filsfils-rtgwg-segment-routing-ldp-interop-00] documents SR co-existence and interworking with other MPLS signaling protocols, if present in the network during a migration, or in case of non-homogeneous deployments.

The operator only needs to allocate one node segment per PE and the SR IGP control-plane automatically builds the required MPLS forwarding constructs from any PE to any PE.

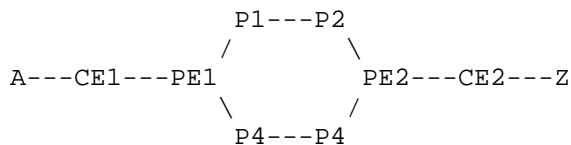


Figure 1: IGP-based MPLS Tunneling

In Figure 1 above, the four nodes A, CE1, CE2 and Z are part of the same VPN.

PE2 advertises (in the IGP) a host address 192.0.2.2/32 with its attached node segment 102.

CE2 advertises to PE2 a route to Z. PE2 binds a local label LZ to that route and propagates the route and its label via MPBGP to PE1 with nhop 192.0.2.2 (PE2 loopback address).

PE1 installs the VPN prefix Z in the appropriate VRF and resolves the next-hop onto the node segment 102. Upon receiving a packet from A destined to Z, PE1 pushes two labels onto the packet: the top label is 102, the bottom label is LZ. 102 identifies the node segment to PE2 and hence transports the packet along the ECMP-aware shortest-path to PE2. PE2 then processes the VPN label LZ and forwards the packet to CE2.

Supporting MPLS services (VPN, VPLS, VPWS) with SR has the following benefits:

- Simple operation: one single intra-domain protocol to operate: the IGP. No need to support IGP synchronization extensions as described in [RFC5443] and [RFC6138].

- Excellent scaling: one Node-SID per PE.

3. MPLS Instantiation of Segment Routing

MPLS instantiation of Segment Routing fits in the MPLS architecture as defined in [RFC3031] both from a control plane and forwarding plane perspective:

- o From a control plane perspective [RFC3031] does not mandate a single signaling protocol. Segment Routing proposes to use the Link State IGP as its use of information flooding fits very well with label stacking on ingress.
- o From a forwarding plane perspective, Segment Routing does not require any change to the forwarding plane.

When applied to MPLS, a Segment is a LSP and the 20 right-most bits of the segment are encoded as a label. This implies that, in the MPLS instantiation, the SID values are allocated within a reduced 20-bit space out of the 32-bit SID space.

The notion of indexed global segment fits the MPLS architecture [RFC3031] as the absolute value allocated to any segment (global or

local) can be managed by a local allocation process (similarly to other MPLS signaling protocols).

If present, SR can coexist and interwork with LDP and RSVP [draft-filsfils-rtgwg-segment-routing-ldp-interop-00].

The source routing model described in [I-D.filsfils-rtgwg-segment-routing] is inherited from the ones proposed by [RFC1940] and [RFC2460]. The source routing model offers the support for explicit routing capability.

Contrary to RSVP-based explicit routes where tunnel midpoints maintain states, SR-based explicit routes only require per-flow states at the ingress edge router where the traffic engineer policy is applied.

Contrary to RSVP-based explicit routes which consist in non-ECMP circuits (similar to ATM/FR), SR-based explicit routes can be built as list of ECMP-aware node segments and hence ECMP-aware traffic engineering is natively supported by SR.

When Segment Routing is instantiated over the MPLS data plane the following applies:

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

- The active segment is the top label.

- The CONTINUE operation is implemented as an MPLS swap operation. When the same SRGB block is used throughout the SR domain, the outgoing label value is equal to the incoming label value . Else, the outgoing label value is [SRGB(next_hop)+index]

- The NEXT operation is implemented as an MPLS pop operation.

- The PUSH operation is implemented as an MPLS push of a label stack.

In conclusion, there are no changes in the operations of the data-plane currently used in MPLS networks.

4. IGP Segments Examples

Assuming the network diagram of Figure 2 and the IP address and IGP Segment allocation of Figure 3, the following examples can be constructed.

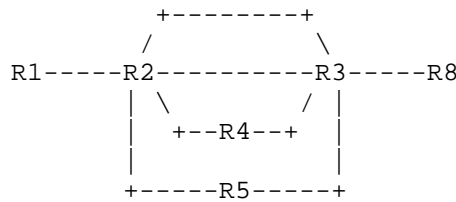


Figure 2: IGP Segments - Illustration

```

+-----+
| IP address allocated by the operator:
|       192.0.2.1/32 as a loopback of R1
|       192.0.2.2/32 as a loopback of R2
|       192.0.2.3/32 as a loopback of R3
|       192.0.2.4/32 as a loopback of R4
|       192.0.2.5/32 as a loopback of R5
|       192.0.2.8/32 as a loopback of R8
|       198.51.100.9/32 as an anycast loopback of R4
|       198.51.100.9/32 as an anycast loopback of R5
|
| SRGB defined by the operator as 1000-5000
|
| Global IGP SID allocated by the operator:
|       1001 allocated to 192.0.2.1/32
|       1002 allocated to 192.0.2.2/32
|       1003 allocated to 192.0.2.3/32
|       1004 allocated to 192.0.2.4/32
|       1008 allocated to 192.0.2.8/32
|       2009 allocated to 198.51.100.9/32
|
| Local IGP SID allocated dynamically by R2
|       for its "north" adjacency to R3: 9001
|       for its "north" adjacency to R3: 9003
|       for its "south" adjacency to R3: 9002
|       for its "south" adjacency to R3: 9003
+-----+

```

Figure 3: IGP Address and Segment Allocation - Illustration

4.1. Example 1

R1 may send a packet P1 to R8 simply by pushing an SR header with segment list {1008}.

1008 is a global IGP segment attached to the IP prefix 192.0.2.8/32. Its semantic is global within the IGP domain: any router forwards a

packet received with active segment 1008 to the next-hop along the ECMP-aware shortest-path to the related prefix.

In conclusion, the path followed by P1 is R1-R2--R3-R8. The ECMP-awareness ensures that the traffic be load-shared between any ECMP path, in this case the two north and south links between R2 and R3.

4.2. Example 2

R1 may send a packet P2 to R8 by pushing an SR header with segment list {1002, 9001, 1008}.

1002 is a global IGP segment attached to the IP prefix 192.0.2.2/32. Its semantic is global within the IGP domain: any router forwards a packet received with active segment 1002 to the next-hop along the shortest-path to the related prefix.

9001 is a local IGP segment attached by node R2 to its north link to R3. Its semantic is local to node R2: R2 switches a packet received with active segment 9001 towards the north link to R3.

In conclusion, the path followed by P2 is R1-R2-north-link-R3-R8.

4.3. Example 3

R1 may send a packet P3 along the same exact path as P1 using a different segment list {1002, 9003, 1008}.

9003 is a local IGP segment attached by node R2 to both its north and south links to R3. Its semantic is local to node R2: R2 switches a packet received with active segment 9003 towards either the north or south links to R3 (e.g. per-flow loadbalancing decision).

In conclusion, the path followed by P3 is R1-R2-any-link-R3-R8.

4.4. Example 4

R1 may send a packet P4 to R8 while avoiding the links between R2 and R3 by pushing an SR header with segment list {1004, 1008}.

1004 is a global IGP segment attached to the IP prefix 192.0.2.4/32. Its semantic is global within the IGP domain: any router forwards a packet received with active segment 1004 to the next-hop along the shortest-path to the related prefix.

In conclusion, the path followed by P4 is R1-R2-R4-R3-R8.

4.5. Example 5

R1 may send a packet P5 to R8 while avoiding the links between R2 and R3 while still benefitting from all the remaining shortest paths (via R4 and R5) by pushing an SR header with segment list {2009, 1008}.

2009 is a global IGP segment attached to the anycast IP prefix 198.51.100.9/32. Its semantic is global within the IGP domain: any router forwards a packet received with active segment 2009 to the next-hop along the shortest-path to the related prefix.

In conclusion, the path followed by P5 is either R1-R2-R4-R3-R8 or R1-R2-R5-R3-R8 .

5. Other Examples of MPLS Segments

In addition to the IGP segments previously described, the SPRING source routing policy applied to MPLS can include MPLS LSP's signaled by LDP, RSVPTE and BGP. The list of examples is non exhaustive. Other form of segments combination can be instantiated through Segment Routing (e.g.: RSVP LSPs combined with LDP or IGP or BGP LSPs).

5.1. LDP LSP segment combined with IGP segments

The example illustrates a segment-routing policy including IGP segments and LDP LSP segments.

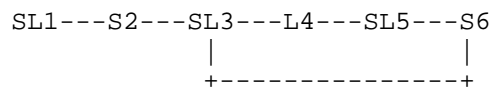


Figure 4: LDP LSP segment combined with IGP segments

We assume that:

- o All links have an IGP cost of 1 except SL3-S6 link which has cost 2.
- o All nodes are in the same IGP area.
- o Nodes SL1, S2, SL3, SL5 and S6 are IGP-SR capable.
- o SL3 and S6 have, respectively, index 3 and 6 assigned to them.
- o All SR nodes have the same SRGB consisting of: [1000, 1999]
- o SL1, SL3, L4 and SL5 are LDP capable.

- o SL1 has a directed LDP session with SL3 and is able to retrieve the SL3 local LDP mapping for FEC SL5: 35
- o The following source-routed policy is defined in S1 for the traffic destined to S6: use path SL1-S2-SL3-L4-SL5-S6 (instead of shortest-path SL1-S2-SL3-S6).

This is realized by programming the following segment-routing policy at S1: for traffic destined to S6, push the ordered segment list: {1003, 35, 1006}, where:

- o 1003 gets the packets from S1 to SL3 via S2.
- o 35 gets the packets from SL3 to SL5 via L4.
- o 1006 gets the packets from SL5 to S6.

The above allows to steer the traffic into path SL1-S2-SL3-L4-SL5-S6 instead of the shortest path SL1-S2-SL3-S6.

5.2. RSVP-TE LSP segment combined with IGP segments

The example illustrates a segment-routing policy including IGP segments and RSVP-TE LSP segments.

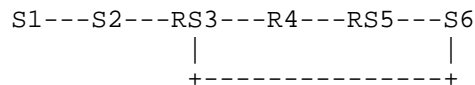


Figure 5: RSVP-TE LSP segment combined with IGP segments

We assume that:

- o All links have an IGP cost of 1 except link RS3-S6 which has cost 2.
- o All nodes are IGP-SR capable except R4.
- o RS3 and R6 have, respectively, index 3 and 6 assigned to them.
- o All SR nodes have the same SRGB consisting of: [1000, 1999]
- o RS3, R4 and RS5 are RSVP-TE capable.
- o An RSVP-TE LSP has been provisioned from RS3 to RS5 via R4.
- o RS3 allocates a binding SID (with value of 135) for this RSVP-TE LSP and signals it in the igp.

- o The following source-routed policy is defined at S1 for the traffic destined to S6: use path S1-S2-RS3-R4-RS5-S6 instead of shortest-path S1-S2-RS3-S6.

This is realized by programming the following segment-routing policy at S1: - for traffic destined to S6, push the ordered segment list: {1003, 135, 1006}, where:

- o 1003 gets the packets from S1 to RS3 via S2.
- o 135 gets the packets from RS3 into the RSVP-TE LSP to RS5 via R4.
- o 1006 gets the packets from RS5 to S6.

The above allows to steer the traffic into path S1-S2-RS3-R4-RS5-S6 instead of the shortest path S1-S2-RS3-S6.

6. Segment List History

In the abstract SR routing model [I-D.filsfils-rtgwg-segment-routing], any node N along the journey of the packet is able to determine where the packet P entered the SR domain and where it will exit. The intermediate node is also able to determine the paths from the ingress edge router to itself, and from itself to the egress edge router.

In the MPLS instantiation, as the packet travels through the SR domain, the stack is depleted and the segment list history is gradually lost.

Future version of this document will describe how this information can be preserved in MPLS domains.

7. IANA Considerations

TBD

8. Manageability Considerations

TBD

9. Security Considerations

TBD

10. Acknowledgements

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

11.2. Informative References

- [I-D.filsfils-rtgwg-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.
- [I-D.filsfils-rtgwg-segment-routing-use-cases]
Filsfils, C., Francois, P., Previdi, S., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., Kini, S., and E. Crabbe, "Segment Routing Use Cases", draft-filsfils-rtgwg-segment-routing-use-cases-02 (work in progress), October 2013.
- [I-D.previdi-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-previdi-isis-segment-routing-extensions-05 (work in progress), February 2014.
- [I-D.psenak-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., and W. Henderickx, "OSPF Extensions for Segment Routing", draft-psenak-ospf-segment-routing-extensions-04 (work in progress), February 2014.

- [I-D.psenak-ospf-segment-routing-ospfv3-extension]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,
Shakir, R., and W. Henderickx, "OSPFv3 Extensions for
Segment Routing", draft-psenak-ospf-segment-routing-
ospfv3-extension-01 (work in progress), February 2014.
- [RFC1940] Estrin, D., Li, T., Rekhter, Y., Varadhan, K., and D.
Zappala, "Source Demand Routing: Packet Format and
Forwarding Specification (Version 1)", RFC 1940, May 1996.
- [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.
- [draft-filsfils-rtgwg-segment-routing-ldp-interop-00]
Filsfils, C. and S. Previdi, "Segment Routing
interoperability with LDP", October 2013.

Authors' Addresses

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

Email: cfilsfil@cisco.com

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

Email: sprevidi@cisco.com

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

Email: bashandy@cisco.com

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange
FR

Email: stephane.litkowski@orange.com

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

Email: Martin.Horneffer@telekom.de

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

Edward Crabbe
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: edc@google.com

spring
Internet-Draft
Intended status: Informational
Expires: August 9, 2014

R. Geib, Ed.
Deutsche Telekom
C. Filsfils
Cisco Systems, Inc.
February 5, 2014

Use case for a scalable and topology aware MPLS data plane monitoring
system
draft-geib-spring-oam-usecase-01

Abstract

This document describes features and a use case of a path monitoring system. Segment based routing enables a scalable and simple method to monitor data plane liveliness of the complete set of paths belonging to a single domain.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 August 9, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. A topology aware MPLS path monitoring system	4
3. SR based OAM use case illustration	5
3.1. Use-case 1 - LSP dataplane liveliness measurement	5
3.2. Use-case 2 - Monitoring a remote bundle	7
3.3. Use-Case 3 - Fault localization	7
4. Applying SR to monitor LDP paths	8
5. PMS monitoring of different Segment ID types	8
6. IANA Considerations	8
7. Security Considerations	8
8. References	9
8.1. Normative References	9
8.2. Informative References	9
Authors' Addresses	9

1. Introduction

It is essential for a network operator to monitor all the forwarding paths observed by the transported user packets. The monitoring flow must be forwarded in dataplane in a similar way as user packets. Problem localization is required.

This document describes a solution to this problem statement and illustrates it with use-cases.

The solution is described for a single IGP MPLS domain.

The solution applies to monitoring of LDP LSP's as well as to monitoring of Segment Routed LSP's. Segment Routing simplifies the solution by the use of IGP-based signalled segments as specified by [ID.sr-isis].

This document adopts the terminology and framework described in [ID.sr-archi]. It further adopts the editorial simplification explained in section 1.2 of the segment routing use-cases [ID.sr-use].

The proposed solution offers several benefits for network monitoring. A single monitoring device is able to monitor the complete set of a domains forwarding paths with OAM packets that never leave data plane. Faults can be localized:

- o by IGP LSA analysis.
- o by correlation between different probes.
- o by MPLS traceroute and adapted ping messages.

The proposed solution requires topology awareness as well as a suitable security architecture. Topology awareness is an essential part of link state IGPs. Adding MPLS topology awareness to an IGP speaking device hence enables a simple and scaleable data plane monitoring mechanism.

MPLS OAM offers flexible features to recognise and execute data paths of an MPLS domain. By utilising the ECMP related tool set of RFC 4379 [RFC4379], a segment based routing LSP monitoring system may:

- o easily detect ECMP functionality and properties of paths at data level.
- o construct monitoring packets executing desired paths also if ECMP is present.

- o limit the MPLS label stack of an OAM packet to a minimum of 3 labels.

MPLS OAM supports detection and execution of ECMP paths quite smart. This document is focused on MPLS path monitoring.

The MPLS path monitoring system described by this document can be realised with pre-Segment based Routing (SR) technology. Making monitoring system aware of a domain's complete MPLS topology from utilising stale MPLS label information, IGP must be monitored and MPLS topology must be timely aligned with IGP topology. Obviously, enhancing IGPs to exchange of MPLS topology information significantly simplifies and stabilises such an MPLS path monitoring system. In addition to IGP extensions, also RFC 4379 may have to be extended to support detection of SR routed paths.

Note that the MPLS path monitoring system may be a specialised system residing at a single interface of the domain to be monitored. As long as measurement packets return to this or another well specified interface, the MPLS monitoring system is the single entity pushing monitoring packet label stacks. Concerns about router label stack pushing capabilities don't apply in this case.

First drafts discussing requirements, extensions of RFC4379 and possible solutions to allow SR usage as described by this document are at hand, see [ID.sr-4379ext] and [ID.sr-oam_detect].

2. A topology aware MPLS path monitoring system

An MPLS path monitoring system (PMS) which is able to learn the IGP LSDB (including the SID's) is able to build a measurement packet which executes any arbitrary chain of paths. Such a monitoring system is topology aware (all related IP addresses, MPLS SIDs and labels).

Let us describe how the PMS can check the liveness of the MPLS transport path between LER i and LER j.

The PMS may do so by sending packets carrying the following minimum address information:

- o Top Label: a path from PMS to LER i This is expressed as Node SID of LER i.
- o Next Label: the path that needs to be monitored from LER i to LER j. If this path is a single physical interface (or a bundle of connected interfaces), it can be expressed by the related AdjSID.

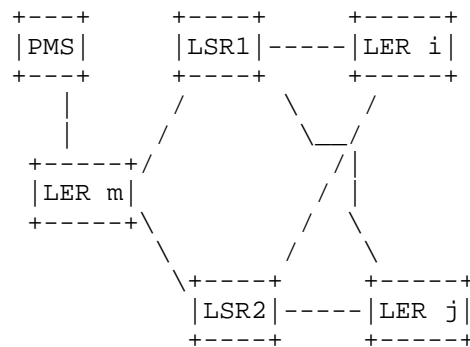
If the shortest path from LER i to LER j is supposed to be monitored, the Node-SID (LER j) can be used. Another option is to insert a list of segments expressing the desired path (hop by hop as an extreme case). If LER i pushes a stack of Labels based on a SR policy decision and this stack of LSPs is to be monitored, the PMS needs an interface to collect the information enabling it to address this SR created path.

- o Next Label or address: the path back to the PMS. Likely, no further segment/label is required here. Indeed, once the packet reaches LER j, the 'steering' part of the solution is done and the probe just needs to return to the PMS. This is best achieved by popping the MPLS stack and revealing a probe packet with PMS as destination address (note that in this case, the source and destination addresses could be the same). In this case, a no SID/label may be assigned to the PMS (if it is a host/server residing in an IP subnet outside the MPLS domain).

Note: if the PMS is an IP host not connected to the MPLS domain, the PMS can send its probe with the list of SIDs/Labels onto a suitable tunnel providing an MPLS access to a router which is part of the monitored MPLS domain.

3. SR based OAM use case illustration

3.1. Use-case 1 - LSP dataplane liveness measurement



Example of a PMS based LSP dataplane liveness measurement

Figure 1

For the sake of simplicity, let's assume that all the nodes are configured with the same SRGB [ID.sr-archi]. as described by section

1.2 of [ID.sr-use].

Let's assign the following Node SIDs to the nodes of the figure: PMS = 10, LER i = 20, LER j = 30.

The aim is to check liveness of the path LER i to LER j. The PMS does this by creating a measurement packet with the following label stack (top to bottom): 20 - 30 - 10.

LER m forwards the packet received from the PMS to LSR1. Assuming Pen-ultimate Hop Popping to be deployed, LSR1 pops the top label and forwards the packet to LER i. There the top label has a value 30 and LER i forwards it to LER j. This will be done transmitting the packet via LSR1 or LSR2. The LSR will again pop the top label. LER j will forward the packet now carrying the top label 10 to the PMS (and it will pass a LSR and LER m).

A few observations on the example:

- o The path PMS to LER i must be stable and it must be detectable.
- o If ECMP is deployed, it may be desired to measure along both possible paths, a packet may use between LER i and LER j. This may be done by using MPLS OAM coded measurement packets with suitable IP destination addresses.
- o The path LER j to PMS to must be stable and it must be detectable.

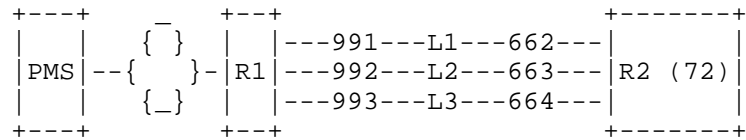
To ensure reliable results, the PMS should be aware of any changes in IGP or MPLS topology.

Determining a path to be executed prior to a measurement may also be done by setting up a label including all node SIDs along that path (if LER1 has Node SID 40 in the example and it should be passed between LER i and LER j, the label stack is 20 - 40 - 30 - 10).

Obviously, the PMS is able to check and monitor data plane liveness of all LSPs in the domain. The PMS may be a router, but could also be dedicated monitoring system. If measurement system reliability is an issue, more than a single PMS may be connected to the MPLS domain.

Monitoring an MPLS domain by a PMS based on SR offers the option of monitoring complete MPLS domains with little effort and very excellent scalability.

3.2. Use-case 2 - Monitoring a remote bundle



SR based probing of all the links of a remote bundle

Figure 2

R1 addresses Lx by the Adjacency SID 99x, while R2 addresses Lx by the Adjacency SID 66(x+1).

In the above figure, the PMS needs to assess the dataplane availability of all the links within a remote bundle connected to routers R1 and R2.

The monitoring system retrieves the SID/Label information from the IGP LSDB and appends the following segment list/label stack: {72, 662, 992, 664} on its IP probe (whose source and destination addresses are the address of the PMS).

MS sends the probe to its connected router. If the connected router is not SR compliant, a tunneling technique can be used to tunnel the probe and its MPLS stack to the first SR router. The MPLS/SR domain then forwards the probe to R2 (72 is the Node SID of R2). R2 forwards the probe to R1 over link L1 (Adjacency SID 662). R1 forwards the probe to R2 over link L2 (Adjacency SID 992). R2 forwards the probe to R1 over link L3 (Adjacency SID 664). R1 then forwards the IP probe to PMS as per classic IP forwarding.

3.3. Use-Case 3 - Fault localization

In the previous example, a uni-directional fault on the middle link from R1 to R2 would be localized by sending the following two probes with respective segment lists:

- o 72, 662, 992, 664
- o 72, 663, 992, 664

The first probe would fail while the second would succeed. Correlation of the measurements reveals that the only difference is

using the Adjacency SID 662 of the middle link from R1 to R2 in the non successful measurement. Assuming the second probe has been routed correctly, the fault must have been occurring in R2 which didn't forward the packet to the interface identified by its Adjacency SID 662.

4. Applying SR to monitor LDP paths

A SR based PMS connected to a MPLS domain consisting of LER and LSR supporting SR and LDP in parallel in all nodes may use SR paths to transmit packets to and from start and end points of LDP paths to be monitored. In the above example, the label stack top to bottom may be as follows, when sent by the PMS:

- o Top: SR based Node-SID of LER i at LER m.
- o Next: LDP label identifying the path to LER j at LER i.
- o Bottom: SR based Node-SID identifying the path to the PMS at LER j

While the mixed operation shown here still requires the PMS to be aware of the LER LDP-MPLS topology, the PMS may learn the SR MPLS topology by IGP and use this information.

5. PMS monitoring of different Segment ID types

MPLS SR topology awareness should allow the SID to monitor liveness of most types of SIDs (this may not be recommendable if a SID identifies an inter domain interface).

To match control plane information with data plane information, RFC4379 should be enhanced to allow collection of data relevant to check all relevant types of Segment IDs.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

As mentioned in the introduction, a PMS monitoring packet should never leave the domain where it originated. It therefore should never use stale MPLS or IGP routing information. Further, assigning different label ranges for different purposes may be useful. A well

known global service level range may be excluded for utilisation within PMS measurement packets. These ideas shouldn't start a discussion. They rather should point out, that such a discussion is required when SR based OAM mechanisms like a SR are standardised.

8. References

8.1. Normative References

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

8.2. Informative References

[ID.sr-4379ext]
IETF, "Label Switched Path (LSP) Ping/Trace for Segment Routing Networks Using MPLS Dataplane", IETF, <http://datatracker.ietf.org/doc/draft-kumar-mpls-spring-lsp-ping/>, 2013.

[ID.sr-archi]
IETF, "Segment Routing Architecture", IETF, <https://datatracker.ietf.org/doc/draft-filsfils-rtgwg-segment-routing/>, 2013.

[ID.sr-isis]
IETF, "IS-IS Extensions for Segment Routing", IETF, <http://datatracker.ietf.org/doc/draft-previdi-isis-segment-routing-extensions/>, 2013.

[ID.sr-oam_detect]
IETF, "Detecting Multi-Protocol Label Switching (MPLS) Data Plane Failures in Source Routed LSPs", IETF, <http://datatracker.ietf.org/doc/draft-kini-spring-mpls-lsp-ping/>, 2013.

[ID.sr-use]
IETF, "Segment Routing Use Cases", IETF, <http://datatracker.ietf.org/doc/draft-filsfils-rtgwg-segment-routing-use-cases/>, 2013.

Authors' Addresses

Ruediger Geib (editor)
Deutsche Telekom
Heinrich Hertz Str. 3-7
Darmstadt, 64295
Germany

Phone: +49 6151 5812747
Email: Ruediger.Geib@telekom.de

Clarence Filsfils
Cisco Systems, Inc.
Brussels,
Belgium

Phone:
Email: cfilsfil@cisco.com

Spring
Internet-Draft
Intended status: Informational
Expires: January 4, 2015

J. Brzozowski
J. Leddy
Comcast
I. Leung
Rogers Communications
S. Previdi
M. Townsley
C. Martin
C. Filsfils
R. Maglione, Ed.
Cisco Systems
July 3, 2014

IPv6 SPRING Use Cases
draft-ietf-spring-ipv6-use-cases-01

Abstract

Source Packet Routing in Networking (SPRING) architecture leverages the source routing paradigm. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with SPRING header. A segment can represent any instruction, topological or service-based. A segment can have a local semantic to the SPRING node or global within the SPRING domain. SPRING allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node to the SPRING domain.

The objective of this document is to illustrate some use cases that need to be taken into account by the Source Packet Routing in Networking (SPRING) architecture.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. IPv6 SPRING use cases	3
2.1. SPRING in the Home Network	5
2.2. SPRING in the Access Network	6
2.3. SPRING in the Data Center	6
2.4. SPRING in the Content Delivery Networks	7
2.5. SPRING in the Core networks	8
3. Acknowledgements	9
4. IANA Considerations	9
5. Security Considerations	10
6. Informative References	10
Authors' Addresses	11

1. Introduction

Source Packet Routing in Networking (SPRING) architecture leverages the source routing paradigm. An ingress node steers a packet through a controlled set of instructions, called segments, by prepending the packet with SPRING header. A segment can represent any instruction, topological or service-based. A segment can represent a local semantic on the SPRING node, or a global semantic within the SPRING domain. SPRING allows one to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node to the SPRING domain.

The SPRING architecture is described in [I-D.filsfils-rtgwg-segment-routing]. The SPRING control plane is agnostic to the dataplane, thus it can be applied to both MPLS and IPv6. In case of MPLS the (list of) segment identifiers are carried

in the MPLS label stack, while for the IPv6 dataplane, a new type of routing extension header is required.

The details of the new routing extension header are described in [I-D.previdi-6man-segment-routing-header] which also covers the security considerations and the aspects related to the deprecation of the IPv6 Type 0 Routing Header described in [RFC5095].

2. IPv6 SPRING use cases

In today's networks, source routing is typically accomplished by encapsulating IP packets in MPLS LSPs that are signaled via RSVP-TE. Therefore, there are scenarios where it may be possible to run IPv6 on top of MPLS, and as such, the MPLS Segment Routing architecture described in [I-D.filsfils-spring-segment-routing-mpls] could be leveraged to provide SPRING capabilities in an IPv6/MPLS environment.

However, there are other cases and/or specific network segments (such as for example the Home Network, the Data Center, etc.) where MPLS may not be available or deployable for lack of support on network elements or for an operator's design choice. In such scenarios a non-MPLS based solution would be preferred by the network operators of such infrastructures.

In addition there are cases where the operators could have made the design choice to disable IPv4, for ease of management and scale (return to single-stack) or due to an address constraint, for example because they do not possess enough IPv4 addresses resources to number all the endpoints and other network elements on which they desire to run MPLS.

In such scenario the support for MPLS operations on an IPv6-only network would be required. However today's IPv6-only networks are not fully capable of supporting MPLS. There is ongoing work in the MPLS Working Group, described in [I-D.ietf-mpls-ipv6-only-gap]

to identify gaps that must be addressed in order to allow MPLS-related protocols and applications to be used with IPv6-only networks. This is another example of scenario where an IPv6-only solution could represent a valid option to solve the problem and meet operators' requirements.

In addition it is worth to note that in today's MPLS dual-stack networks IPv4 traffic is labeled while IPv6 traffic is usually natively routed, not label-switched. Therefore in order to be able to provide Traffic Engineering "like" capabilities for IPv6 traffic additional/alternative encapsulation mechanisms would be required.

In summary there is a class of use cases that motivate an IPv6 data plane. The authors identify some fundamental scenarios that, when recognized in conjunction, strongly indicate an IPv6 data plane:

1. There is a need or desire to impose source-routing semantics within an application or at the edge of a network (for example, a CPE or home gateway)
2. There is a strict lack of an MPLS dataplane
3. There is a need or desire to remove routing state from any node other than the source, such that the source is the only node that knows and will know the path a packet will take, a priori
4. There is a need to connect millions of addressable segment endpoints, thus high routing scalability is a requirement. IPv6 addresses are inherently summarizable: a very large operator could scale by summarizing IPv6 subnets at various internal boundaries. This is very simple and is a basic property of IP routing. MPLS node segments are not summarizable. To reach the same scale, an operator would need to introduce additional complexity, such as mechanisms described in [I-D.ietf-mpls-seamless-mpls]

In any environment with requirements such as those listed above, an IPv6 data plane provides a powerful combination of capabilities for a network operator to realize benefits in explicit routing, protection and restoration, high routing scalability, traffic engineering, service chaining, service differentiation and application flexibility via programmability.

This section will describe some scenarios where MPLS may not be present and it will highlight how the SPRING architecture could be used to address such use cases, particularly, when an MPLS data plane is neither present nor desired.

The use cases described in the section do not constitute an exhaustive list of all the possible scenarios; this section only includes some of the most common envisioned deployment models for IPv6 Segment Routing.

In addition to the use cases described in this document the SPRING architecture can be applied to all the use cases described in [I-D.filsfils-rtgwg-segment-routing-use-cases] for the SPRING MPLS data plane, when an IPv6 data plane is present.

2.1. SPRING in the Home Network

An IPv6-enabled home network provides ample globally routed IP addresses for all devices in the home. An IPv6 home network with multiple egress points and associated provider-assigned prefixes will, in turn, provide multiple IPv6 addresses to hosts. A homenet performing Source and Destination Routing ([I-D.troan-homenet-sadr]) will ensure that packets exit the home at the appropriate egress based on the associated delegated prefix for that link.

A SPRING enabled home provides the possibility for imposition of a Segment List by end-hosts in the home, or a customer edge router in the home. If the Segment List is enabled at the customer edge router, that router is responsible for classifying traffic and inserting the appropriate Segment List. If hosts in the home have explicit source selection rules (see [I-D.lepape-6man-prefix-metadata]), classification can be based on source address or associated network egress point, avoiding the need for DPI-based implicit classification techniques. If the Segment List is inserted by the host itself, it is important to know which networks can interpret the SPRING header. This information can be provided as part of host configuration as a property of the configured IP address (see [I-D.bhandari-dhc-class-based-prefix]).

The ability to steer traffic to an appropriate egress or utilize a specific type of media (e.g., low-power, WIFI, wired, femto-cell, bluetooth, MOCA, HomePlug, etc.) within the home itself are obvious cases which may be of interest to an application running within a home network.

Steering to a specific egress point may be useful for a number of reasons, including:

- o Regulatory
- o Performance of a particular service associated with a particular link
- o Cost imposed due to data-caps or per-byte charges
- o Home vs. work traffic in homes with one or more teleworkers, etc.
- o Specific services provided by one ISP vs. another

Information included in the Segment List, whether imposed by the end-host itself, a customer edge router, or within the access network of the ISP, may be of use at the far ends of the data communication as well. For example, an application running on an end-host with

application-support in a data center can utilize the Segment List as a channel to include information that affects its treatment within the data center itself, allowing for application-level steering and load-balancing without relying upon implicit application classification techniques at the data-center edge. Further, as more and more application traffic is encrypted, the ability to extract (and include in the Segment List) just enough information to enable the network and data center to load-balance and steer traffic appropriately becomes more and more important.

2.2. SPRING in the Access Network

Access networks deliver a variety of types of traffic from the service provider's network to the home environment and from the home towards the service provider's network.

For bandwidth management or related purposes, the service provider may want to associate certain types of traffic to specific physical or logical downstream capacity pipes.

This mapping is not the same thing as classification and scheduling. In the Cable access network, each of these pipes are represented at the DOCSIS layer as different service flows, which are better identified as differing data links. As such, creating this separation allows an operator to differentiate between different types of content and perform a variety of differing functions on these pipes, such as egress vectoring, byte capping, regulatory compliance functions, and billing.

In a cable operator's environment, these downstream pipes could be a specific QAM, a DOCSIS service flow or a service group.

Similarly, the operator may want to map traffic from the home sent towards the service provider's network to specific upstream capacity pipes. Information carried in a packet's SPRING header could provide the target pipe for this specific packet. The access device would not need to know specific details about the packet to perform this mapping; instead the access device would only need to know how to map the SR SID value to the target pipe.

2.3. SPRING in the Data Center

A key use case for SPRING is to cause a packet to follow a specific path through the network. One can think of the service function performed at each SPRING node to be forwarding. More complex service functions could be applied to the packet by a SPRING node including accounting, IDS, load balancing, and fire walling.

The term "Service Function Chain", as defined in [I-D.ietf-sfc-problem-statement], it is used to describe an ordered set of service functions that must be applied to packets.

A service provider may choose to have these service functions performed external to the routing infrastructure, specifically on either dedicated physical servers or within VMs running on a virtualization platform.

[I-D.kumar-sfc-dc-use-cases] describes use cases that demonstrate the applicability of Service Function Chaining (SFC) within a data center environment and provides SFC requirements for data center centric use cases.

2.4. SPRING in the Content Delivery Networks

The rise of online video applications and new, video-capable IP devices has led to an explosion of video traffic traversing network operator infrastructures. In the drive to reduce the capital and operational impact of the massive influx of online video traffic, as well as to extend traditional TV services to new devices and screens, network operators are increasingly turning to Content Delivery Networks (CDNs).

Several studies showed the benefits of connecting caches in a hierarchical structure following the hierarchical nature of the Internet. In a cache hierarchy one cache establishes peering relationships with its neighbor caches. There are two types of relationship: parent and sibling. A parent cache is essentially one level up in a cache hierarchy. A sibling cache is on the same level. Multiple levels of hierarchy are commonly used in order to build efficient caches architecture.

In an environment, where each single cache system can be uniquely identified by its own IPv6 address, a Segment List containing a sequence of the caches in a hierarchy can be built. At each node (cache) present in the Segment List a TCP session to port 80 is established and if the requested content is found at the cache (cache hits scenario) the sequence ends, even if there are more nodes in the list.

To achieve the behavior described above, in addition to the Segment List, which specifies the path to be followed to explore the hierarchic architecture, a way to instruct the node to take a specific action is required. The function to be performed by a service node can be carried into a new header called Network Service Header (NSH) defined in [I-D.quinn-sfc-nsh]. A Network Service Header (NSH) is metadata added to a packet that is used to create a

service plane. The service header is added by a service classification function that determines which packets require servicing, and correspondingly which service path to follow to apply the appropriate service.

In the above example the service to be performed by the service node was to establish a TCP session to port 80, but in other scenarios different functions may be required. Another example of action to be taken by the service node is the capability to perform transformations on payload data, like real-time video transcode option (for rate and/or resolution).

The use of SPRING together with the NSH allows building flexible service chains where the topological information related to the path to be followed is carried into the Segment List while the "service plane related information" (function/action to be performed) is encoded in the metadata, carried into the NSH. The details about using SPRING together with NSH will be described in a separate document.

2.5. SPRING in the Core networks

MPLS is a well-known technology widely deployed in many IP core networks. However there are some operators that do not run MPLS everywhere in their core network today, thus moving forward they would prefer to have an IPv6 native infrastructure for the core network.

While the overall amount of traffic offered to the network continues to grow and considering that multiple types of traffic with different characteristics and requirements are quickly converging over single network architecture, the network operators are starting to face new challenges.

Some operators are looking at the possibility to setup an explicit path based on the IPv6 source address for specific types of traffic in order to efficiently use their network infrastructure. In case of IPv6 some operators are currently assigning or plan to assign IPv6 prefix(es) to their IPv6 customers based on regions/geography, thus the subscriber's IPv6 prefix could be used to identify the region where the customer is located. In such environment the IPv6 source address could be used by the Edge nodes of the network to steer traffic and forward it through a specific path other than the optimal path.

The need to setup a source-based path, going through some specific middle/intermediate points in the network may be related to different requirements:

- o The operator may want to be able to use some high bandwidth links for specific type of traffic (like video) avoiding the need for over-dimensioning all the links of the network;
- o The operator may want to be able to setup a specific path for delay sensitive applications;
- o The operator may have the need to be able to select one (or multiple) specific exit point(s) at peering points when different peering points are available;
- o The operator may have the need to be able to setup a source based path for specific services in order to be able to reach some servers hosted in some facilities not always reachable through the optimal path;
- o The operator may have the need to be able to provision guaranteed disjoint paths (so-called dual-plane network) for diversity purposes

All these scenarios would require a form of traffic engineering capabilities in IP core networks not running MPLS and not willing to run it.

IPv4 protocol does not provide such functionalities today and it is not the intent of this document to address the IPv4 scenario, both because this may create a lot of backward compatibility issues with currently deployed networks and for the security issues that may raise.

The described use cases could be addressed with the SPRING architecture by having the Edge nodes of network to impose a Segment List on specific traffic flows, based on certain classification criteria that would include source IPv6 address.

3. Acknowledgements

The authors would like to thank Brian Field, Robert Raszuk, Wes George, John G. Scudder and Yakov Rekhter for their valuable comments and inputs to this document.

4. IANA Considerations

This document does not require any action from IANA.

5. Security Considerations

There are a number of security concerns with source routing at the IP layer [RFC5095]. The new IPv6-based routing header will be defined in way that blind attacks are never possible, i.e., attackers will be unable to send source routed packets that get successfully processed, without being part of the negotiations for setting up the source routes or being able to eavesdrop legitimate source routed packets. In some networks this base level security may be complemented with other mechanisms, such as packet filtering, cryptographic security, etc.

6. Informative References

[I-D.bhandari-dhc-class-based-prefix]

Systems, C., Halwasia, G., Gundavelli, S., Deng, H., Thiebaut, L., Korhonen, J., and I. Farrer, "DHCPv6 class based prefix", draft-bhandari-dhc-class-based-prefix-05 (work in progress), July 2013.

[I-D.filsfils-rtgwg-segment-routing]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.

[I-D.filsfils-rtgwg-segment-routing-use-cases]

Filsfils, C., Francois, P., Previdi, S., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., Kini, S., and E. Crabbe, "Segment Routing Use Cases", draft-filsfils-rtgwg-segment-routing-use-cases-02 (work in progress), October 2013.

[I-D.filsfils-spring-segment-routing-mpls]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-filsfils-spring-segment-routing-mpls-02 (work in progress), June 2014.

[I-D.ietf-mpls-ipv6-only-gap]

George, W. and C. Pignataro, "Gap Analysis for Operating IPv6-only MPLS Networks", draft-ietf-mpls-ipv6-only-gap-00 (work in progress), April 2014.

- [I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.
- [I-D.ietf-sfc-problem-statement]
Quinn, P. and T. Nadeau, "Service Function Chaining Problem Statement", draft-ietf-sfc-problem-statement-07 (work in progress), June 2014.
- [I-D.kumar-sfc-dc-use-cases]
Surendra, S., Obediente, C., Tufail, M., Majee, S., and C. Captari, "Service Function Chaining Use Cases In Data Centers", draft-kumar-sfc-dc-use-cases-02 (work in progress), May 2014.
- [I-D.lepape-6man-prefix-metadata]
Pape, M., Systems, C., and I. Farrer, "IPv6 Prefix Metadata and Usage", draft-lepape-6man-prefix-metadata-00 (work in progress), July 2013.
- [I-D.previdi-6man-segment-routing-header]
Previdi, S., Filsfils, C., Field, B., and I. Leung, "IPv6 Segment Routing Header (SRH)", draft-previdi-6man-segment-routing-header-01 (work in progress), June 2014.
- [I-D.quinn-sfc-nsh]
Quinn, P., Guichard, J., Fernando, R., Surendra, S., Smith, M., Yadav, N., Agarwal, P., Manur, R., Chauhan, A., Elzur, U., McConnell, B., and C. Wright, "Network Service Header", draft-quinn-sfc-nsh-02 (work in progress), February 2014.
- [I-D.troan-homenet-sadr]
Troan, O. and L. Colitti, "IPv6 Multihoming with Source Address Dependent Routing (SADR)", draft-troan-homenet-sadr-01 (work in progress), September 2013.
- [RFC5095] Abley, J., Savola, P., and G. Neville-Neil, "Deprecation of Type 0 Routing Headers in IPv6", RFC 5095, December 2007.

Authors' Addresses

John Brzozowski
Comcast

Email: john_brzozowski@cable.comcast.com

John Leddy
Comcast

Email: John_Leddy@cable.comcast.com

Ida Leung
Rogers Communications
8200 Dixie Road
Brampton, ON L6T 0C1
CANADA

Email: Ida.Leung@rci.rogers.com

Stefano Previdi
Cisco Systems
Via Del Serafico, 200
Rome 00142
Italy

Email: sprevidi@cisco.com

Mark Townsley
Cisco Systems

Email: townsley@cisco.com

Christian Martin
Cisco Systems

Email: martincj@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
BE

Email: cfilsfil@cisco.com

Roberta Maglione (editor)
Cisco Systems
181 Bay Street
Toronto M5J 2T3
Canada

Email: robmg1@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 28, 2014

S. Previdi, Ed.
C. Filsfils, Ed.
Cisco Systems, Inc.
B. Decraene
S. Litkowski
Orange
M. Horneffer
R. Geib
Deutsche Telekom
R. Shakir
British Telecom
R. Raszuk
Individual
June 26, 2014

SPRING Problem Statement and Requirements
draft-ietf-spring-problem-statement-01

Abstract

The ability for a node to specify a forwarding path, other than the normal shortest path, that a particular packet will traverse, benefits a number of network functions. Source-based routing mechanisms have previously been specified for network protocols, but have not seen widespread adoption. In this context, the term 'source' means 'the point at which the explicit route is imposed'.

This document outlines various use cases, with their requirements, that need to be taken into account by the Source Packet Routing in Networking (SPRING) architecture for unicast traffic. Multicast use-cases and requirements are out of scope of this document.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on December 28, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Dataplanes	4
3. IGP-based MPLS Tunneling	4
3.1. Example of IGP-based MPLS Tunnels	4
4. Fast Reroute	5
5. Traffic Engineering	5
5.1. Examples of Traffic Engineering Use Cases	6
5.1.1. Traffic Engineering without Bandwidth Admission Control	6
5.1.2. Traffic Engineering with Bandwidth Admission Control	10
6. Interoperability with non-SPRING nodes	14
7. OAM	14
8. Security	14
9. IANA Considerations	15
10. Manageability Considerations	15
11. Security Considerations	15
12. Acknowledgements	15
13. References	15
13.1. Normative References	15
13.2. Informative References	15
Authors' Addresses	17

1. Introduction

The ability for a node to specify a unicast forwarding path, other than the normal shortest path, that a particular packet will traverse, benefits a number of network functions, for example:

- Some types of network virtualization, including multi-topology networks and the partitioning of network resources for VPNs

- Network, link, path and node protection such as fast re-route

- Network programmability

- OAM techniques

- Simplification and reduction of network signaling components

- Load balancing and traffic engineering

Source-based routing mechanisms have previously been specified for network protocols, but have not seen widespread adoption other than in MPLS traffic engineering.

These network functions may require greater flexibility and per packet source imposed routing than can be achieved through the use of the previously defined methods. In the context of this charter, 'source' means 'the point at which the explicit route is imposed'.

In this context, Source Packet Routing in Networking (SPRING) architecture is being defined in order to address the use cases and requirements described in this document.

SPRING architecture should allow incremental and selective deployment without any requirement of flag day or massive upgrade of all network elements.

SPRING architecture should allow optimal virtualization: put policy state in the packet header and not in the intermediate nodes along the path. Hence, the policy is completely virtualized away from midpoints and tail-ends.

SPRING architecture objective is not to replace existing source routing and traffic engineering mechanisms but rather complement them and address use cases where removal of signaling and path state in the core is a requirement.

2. Dataplanes

The SPRING architecture should be general in order to ease its applicability to different dataplanes.

MPLS dataplane doesn't require any modification in order to apply a source-based routed model (e.g.: [I-D.filsfils-spring-segment-routing-mpls]).

IPv6 specification [RFC2460], amended by [RFC6564] and [RFC7045], defines the Routing Extension Header which provides IPv6 source-based routing capabilities.

The SPRING architecture should leverage existing MPLS dataplane without any modification and leverage IPv6 dataplane with a new IPv6 Routing Header Type (IPv6 Routing Header is defined in [RFC2460]).

3. IGP-based MPLS Tunneling

The source-based routing model, applied to the MPLS dataplane, offers the ability to tunnel services (VPN, VPLS, VPWS) from an ingress PE to an egress PE, with or without the expression of an explicit path and without requiring forwarding plane or control plane state in intermediate nodes.

The source-based routing model, applied to the MPLS dataplane, offers the ability to tunnel unicast services (VPN, VPLS, VPWS) from an ingress PE to an egress PE, with or without the expression of an explicit path and without requiring forwarding plane or control plane state in intermediate nodes. p2mp and mp2mp tunnels are out of the scope of this document.

3.1. Example of IGP-based MPLS Tunnels

This section illustrates an example use-case taken from [I-D.filsfils-spring-segment-routing-use-cases].

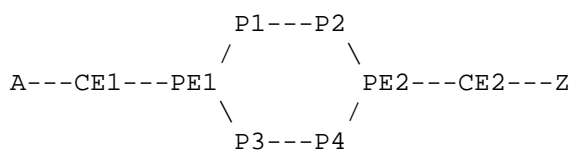


Figure 1: IGP-based MPLS Tunneling

In Figure 1 above, the four nodes A, CE1, CE2 and Z are part of the same VPN. CE2 advertises to PE2 a route to Z. PE2 binds a local label LZ to that route and propagates the route and its label via

MPBGP to PE1 with nhop 192.168.0.2. PE1 installs the VPN prefix Z in the appropriate VRF and resolves the next-hop onto the node segment associated with PE2.

In order to cope with the reality of current deployments, the SPRING architecture should allow PE to PE forwarding according to the IGP shortest path without the addition of any other signaling protocol. The packet each PE forwards across the network will contain (within their label stack) the necessary information derived from the topology database in order to deliver the packet to the remote PE.

4. Fast Reroute

FRR technologies have been deployed by network operators in order to cope with link or node failures through pre-computation of backup paths.

The SPRING architecture should address following requirements:

- o support of FRR on any topology
- o pre-computation and setup of backup path without any additional signaling (other than the regular IGP/BGP protocols)
- o support of shared risk constraints
- o support of node and link protection
- o support of microloop avoidance

Further illustrations of the problem statement for FRR are to be found in [I-D.francois-spring-resiliency-use-case].

5. Traffic Engineering

Traffic Engineering has been addressed using IGP protocol extensions (for resources information propagation) and RSVP-TE for signaling explicit paths. Different contexts and modes have been defined (single vs. multiple domains, with or without bandwidth admission control, centralized vs. distributed path computation, etc).

In all cases, one of the major components of the TE architecture is the soft state based signaling protocol (RSVP-TE) which is used in order to signal and establish the explicit path. Each path, once computed, need to be signaled and state for each path must be present in each node traversed by the path. This incurs a scalability problem especially in the context of SDN where traffic differentiation may be done at a finer granularity (e.g.: application

specific). Also the amount of state needed to be maintained and periodically refreshed in all involved nodes contributes significantly to complexity and the number of failures cases, and thus increases operational effort while decreasing overall network reliability.

The source-based routing model allows traffic engineering to be implemented without the need of a signaling component.

The SPRING architecture should support traffic engineering, including:

- o loose or strict options
- o bandwidth admission control
- o distributed vs. centralized model (PCE, SDN Controller)
- o disjointness in dual-plane networks
- o egress peering traffic engineering
- o load-balancing among non-parallel links
- o Limiting (scalable, preferably zero) per-service state and signaling on midpoint and tail-end routers.
- o ECMP-awareness
- o node resiliency property (i.e.: the traffic-engineering policy is not anchored to a specific core node whose failure could impact the service.

5.1. Examples of Traffic Engineering Use Cases

As documented in [I-D.filsfils-spring-segment-routing-use-cases] here follows the description of two sets of use cases:

- o Traffic Engineering without Admission Control
- o Traffic Engineering with Admission Control

5.1.1. Traffic Engineering without Bandwidth Admission Control

In this section, we describe Traffic Engineering use-cases without bandwidth admission control.

5.1.1.1. Disjointness in dual-plane networks

Many networks are built according to the dual-plane design, as illustrated in Figure 2:

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 and aggregation region K

$C(2,k)$ is part of plane 2 and aggregation region K

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

The core nodes of a given region are directly connected.
Inter-region links only connect core nodes of the same plane.

$\{C(1,k) \text{ has a link to } C(1, j)\}$ iff $\{C(2,k) \text{ has a link to } C(2, j)\}$.

The distribution of these links depends on the topological properties of the core of the AS. The design rule presented above specifies that these links appear in both core planes.

We assume a common design rule found in such deployments: the inter-plane link costs ($C_{ik}-C_{jk}$ where $i \neq j$) are set such that the route to an edge destination from a given plane stays within the plane unless the plane is partitioned.

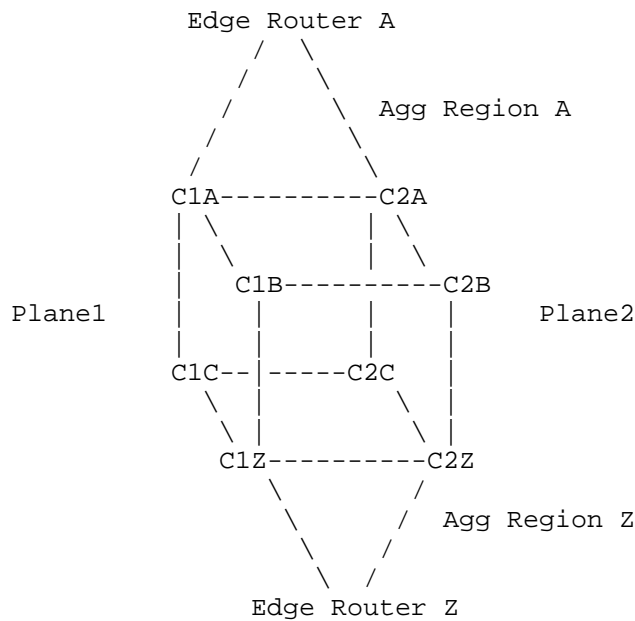


Figure 2: Dual-Plane Network and Disjointness

In this scenario, the operator requires the ability to deploy different strategies. For example, A should be able to use the three following options:

- o the traffic is load-balanced across any ECMP path through the network
- o the traffic is load-balanced across any ECMP path within the Plane1 of the network
- o the traffic is load-balanced across any ECMP path within the Plane2 of the network

Most of the data traffic from A to Z would use the first option, such as to exploit the capacity efficiently. The operator would use the two other choices for specific premium traffic that has requested disjoint transport.

The SPRING architecture should support this use case with the following requirements:

- o Zero per-service state and signaling on midpoint and tail-end routers.

- o ECMP-awareness.
- o Node resiliency property: the traffic-engineering policy is not anchored to a specific core node whose failure could impact the service.

5.1.1.2. Egress Peering Traffic Engineering

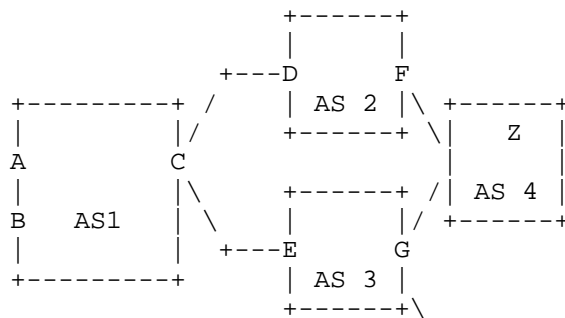


Figure 3: Egress peering traffic engineering

Let us assume, in the network depicted in Figure 3, that:

C in AS1 learns about destination Z of AS 4 via two BGP paths (AS2, AS4) and (AS3, AS4).

C may or may not be configured so to enforce next-hop-self behavior before propagating the paths within AS1.

C may propagate all the paths to Z within AS1 (add-path).

C may install in its FIB only the route via AS2, or only the route via AS3, or both.

In that context, SPRING should allow the operator of AS1 to apply the following traffic-engineering policy, regardless the configured behavior of next-hop-self:

Steer 60% of the Z-destined traffic received at A via AS2 and 40% via AS3.

Steer 80% of the Z-destined traffic received at B via AS2 and 20% via AS3.

While egress routers are known in the routing domain (generally through their loopback address), the SPRING architecture should enable following:

- o identify the egress interfaces of an egress node
- o identify the peering neighbors of an egress node
- o identify the peering ASes of an egress node

With these identifiers known in the domain, the SPRING architecture should allow an ingress node to select the exit point of a packet as any combination of an egress node, an egress interface, a peering neighbor, and a peering AS.

5.1.1.3. Load-balancing among non-parallel links

The SPRING architecture should allow a given node should be able to load share traffic across multiple non parallel links even if these ones lead to different neighbors. This may be useful to support traffic engineering policies.

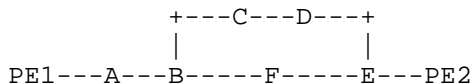


Figure 4: Multiple (non-parallel) Adjacencies

In the above example, the operator requires PE1 to load-balance its PE2-destined traffic between the ABCDE and ABFE paths.

5.1.2. Traffic Engineering with Bandwidth Admission Control

The implementation of bandwidth admission control within a network (and its possible routing consequence which consists in routing along explicit paths where the bandwidth is available) requires a capacity planning process.

The spreading of load among ECMP paths is a key attribute of the capacity planning processes applied to packet-based networks.

5.1.2.1. Capacity Planning Process

Capacity Planning anticipates the routing 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-paths 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 below, and assuming a default IGP metric of 1 and IGP metric of 2 for link GF, 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.

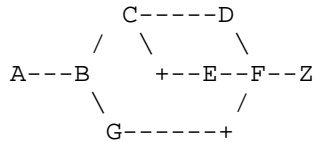


Figure 5: Capacity Planning an ECMP-based demand

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. It is a key efficiency objective to spread the traffic among as many ECMP paths as possible.

This is illustrated in the below network diagram which consists of a subset of a network where already 5 ECMP paths are observed from A to M.

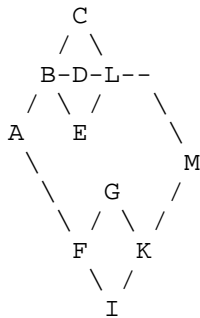


Figure 6: ECMP Topology Example

When the capacity planning process detects that a traffic growth scenario and topology variation would lead to congestion, a capacity increase is triggered and if it cannot be deployed in due time, a traffic engineering solution is activated within the network.

A basic traffic engineering objective 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.

SPRING architecture should offer a simple support for ECMP-based shortest path placement as well as for explicit path policy without incurring additional signaling in the domain. This includes:

- o the ability to steer a packet across a set of ECMP paths
- o the ability to diverge from a set of ECMP shortest paths to one or more paths not in the set of shortest paths

5.1.2.2. SDN/SR use-case

The SDN use-case lies in the SDN controller, (e.g.: Stateful PCE as described in [I-D.ietf-pce-stateful-pce]).

The SDN controller is responsible to control the evolution of the traffic matrix and topology. It accepts or denies the addition of new traffic into the network. It decides how to route the accepted traffic. It monitors the topology and upon topological change, determines the minimum traffic that should be rerouted on an alternate path to alleviate a bandwidth congestion issue.

The algorithms supporting this behavior are a local matter of the SDN controller and are outside the scope of this document.

The means of collecting traffic and topology information are the same as what would be used with other SDN-based traffic-engineering solutions (e.g. [RFC7011] and [I-D.ietf-idr-ls-distribution]).

The means of instantiating policy information at a traffic-engineering head-end are the same as what would be used with other SDN-based traffic-engineering solutions (e.g.: [I-D.ietf-i2rs-architecture], [I-D.crabbe-pce-pce-initiated-lsp] and [I-D.sivabalan-pce-segment-routing]).

In the context of Centralized-Based Optimization and the SDN use-case, here are the benefits that the SPRING architecture should deliver:

Explicit routing capability with or without ECMP-awareness.

No signaling hop-by-hop through the network.

State is only maintained at the policy head-end. No state is maintained at mid-points and tail-ends.

Automated guaranteed FRR for any topology.

Optimum virtualization: the policy state is in the packet header and not in the intermediate nodes along the path. The policy is completely virtualized away from midpoints and tail-ends.

Highly responsive to change: the SDN Controller only needs to apply a policy change at the head-end. No delay is introduced due to programming the midpoints and tail-end along the path.

5.1.2.2.1. SDN Example

The data-set consists in a full-mesh of 12000 explicitly-routed tunnels observed on a real network. These tunnels resulted from distributed headend-based CSPF computation.

We measured that only 65% of the traffic is forwarded over its shortest path.

Three well-known defects are illustrated in this data set:

The lack of ECMP support in explicitly routed tunnels: ATM-alike traffic-steering mechanisms steer the traffic along a non-ECMP path.

The increase of the number of explicitly-routed non-ECMP tunnels to enumerate all the ECMP options.

The inefficiency of distributed optimization: too much traffic is forwarded off its shortest path.

We applied the SDN use-case to this dataset implying a source route model where the path of the packet is encoded within the packet itself. This means that:

The distributed CSPF computation is replaced by centralized optimization and BW admission control, supported by the SDN Controller.

As part of the optimization, we also optimized the IGP-metrics such as to get a maximum of traffic load-spread among ECMP paths by default.

The traffic-engineering policies are supported by a source route model (e.g.: [I-D.filsfils-spring-segment-routing]).

As a result, we measured that 98% of the traffic would be kept on its normal policy (over the shortest-path) and only 2% of the traffic requires a path away from the shortest-path.

Let us highlight a few benefits:

98% of the traffic-engineering head-end policies are eliminated.

Indeed, by default, an ingress edge node capable of injecting source routed packets steers the traffic to the egress edge node. No configuration or policy needs to be maintained at the ingress edge node to realize this.

100% of the states at mid/tail nodes are eliminated.

6. Interoperability with non-SPRING nodes

SPRING must inter-operate with non-SPRING nodes.

An illustration of interoperability between SPRING and other MPLS Signalling Protocols (LDP) is described here in [I-D.filsfils-spring-segment-routing-ldp-interop].

Interoperability with IPv6 non-SPRING nodes will be described in a future document.

7. OAM

The SPRING WG should provide OAM and the management needed to manage SPRING enabled networks. The SPRING procedures may also be used as a tool for OAM in SPRING enabled networks.

OAM use cases and requirements are described in [I-D.geib-spring-oam-usecase] and [I-D.kumar-spring-sr-oam-requirement].

8. Security

There is an assumed trust model such that any node imposing an explicit route on a packet is assumed to be allowed to do so. In such context trust boundaries should strip explicit routes from a packet.

For each data plane technology that SPRING specifies, a security analysis must be provided showing how protection is provided against an attacker disrupting the network by for example, maliciously injecting SPRING packets.

9. IANA Considerations

TBD

10. Manageability Considerations

TBD

11. Security Considerations

TBD

12. Acknowledgements

The authors would like to thank Yakov Rekhter for his contribution to this document.

13. References

13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
- [RFC6564] Krishnan, S., Woodyatt, J., Kline, E., Hoagland, J., and M. Bhatia, "A Uniform Format for IPv6 Extension Headers", RFC 6564, April 2012.
- [RFC7045] Carpenter, B. and S. Jiang, "Transmission and Processing of IPv6 Extension Headers", RFC 7045, December 2013.

13.2. Informative References

- [I-D.crabbe-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-crabbe-pce-pce-initiated-lsp-03 (work in progress), October 2013.
- [I-D.filsfils-spring-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-spring-segment-routing-03 (work in progress), June 2014.

[I-D.filsfils-spring-segment-routing-ldp-interop]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing interoperability with LDP", draft-filsfils-spring-segment-routing-ldp-interop-01 (work in progress), April 2014.

[I-D.filsfils-spring-segment-routing-mpls]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-filsfils-spring-segment-routing-mpls-02 (work in progress), June 2014.

[I-D.filsfils-spring-segment-routing-use-cases]

Filsfils, C., Francois, P., Previdi, S., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., Kini, S., and E. Crabbe, "Segment Routing Use Cases", draft-filsfils-spring-segment-routing-use-cases-00 (work in progress), March 2014.

[I-D.francois-spring-resiliency-use-case]

Francois, P., Filsfils, C., Decraene, B., and R. Shakir, "Use-cases for Resiliency in SPRING", draft-francois-spring-resiliency-use-case-02 (work in progress), April 2014.

[I-D.geib-spring-oam-usecase]

Geib, R. and C. Filsfils, "Use case for a scalable and topology aware MPLS data plane monitoring system", draft-geib-spring-oam-usecase-01 (work in progress), February 2014.

[I-D.ietf-i2rs-architecture]

Atlas, A., Halpern, J., Hares, S., Ward, D., and T. Nadeau, "An Architecture for the Interface to the Routing System", draft-ietf-i2rs-architecture-04 (work in progress), June 2014.

[I-D.ietf-idr-ls-distribution]

Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-05 (work in progress), May 2014.

[I-D.ietf-pce-stateful-pce]

Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-09 (work in progress), June 2014.

[I-D.kumar-spring-sr-oam-requirement]

Kumar, N., Pignataro, C., Akiya, N., Geib, R., and G. Mirsky, "OAM Requirements for Segment Routing Network", draft-kumar-spring-sr-oam-requirement-00 (work in progress), February 2014.

[I-D.sivabalan-pce-segment-routing]

Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and R. Raszuk, "PCEP Extensions for Segment Routing", draft-sivabalan-pce-segment-routing-02 (work in progress), October 2013.

[RFC7011] Claise, B., Trammell, B., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, September 2013.

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

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange
FR

Email: stephane.litkowski@orange.com

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

Email: Martin.Horneffer@telekom.de

Ruediger Geib
Deutsche Telekom
Heinrich Hertz Str. 3-7
Darmstadt 64295
DE

Email: Ruediger.Geib@telekom.de

Rob Shakir
British Telecom
London
UK

Email: rob.shakir@bt.com

Robert Raszuk
Individual

Email: robert@raszuk.net

SPRING WG
Internet-Draft
Intended status: Informational
Expires: December 31, 2014

B. Khasnabish
ZTE TX Inc.
F. Hu
ZTE Corporation
M. Luis
Telefonica I+D
June 29, 2014

Segment Routing in IP RAN use case
draft-kh-spring-ip-ran-use-case-01.txt

Abstract

Segment Routing (SR) leverages the source routing paradigm. An ingress node steers a packet through a controlled set of instructions, called segments, by pre-pending the packet with an SR header. A segment can represent any instruction, topological or service-based. A segment can have a local semantic to an SR node or global within an SR domain. SR allows one to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain. This document introduces the segment routing in IP Radio Access Network (IP RAN, mobile backhaul network) use case. Additional requirements to support segment routing in the IP RAN scenarios are discussed.

Status of This Memo

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

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

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

This Internet-Draft will expire on December 31, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Conventions and Abbreviations	3
3. IP RAN Network Architecture	3
3.1. IP RAN Network Scenario	3
3.2. Requirements for IP RAN network	4
4. Benefit for segment routing in IP RAN network	5
5. Unified Service Deployment	6
5.1. Requirement for Control Node	8
5.2. Requirement for Forwarding Node	8
5.2.1. Forwarding Node Structure	8
6. Security Considerations	9
7. Acknowledgements	9
8. IANA Considerations	9
9. Normative References	9
Authors' Addresses	10

1. Introduction

Segment Routing (SR) leverages the source routing paradigm. An ingress node steers a packet through a controlled set of instructions, called segments, by pre-pending the packet with an SR header. A segment can represent any instruction, topological or service-based. A segment can have a local semantic to an SR node or global within an SR domain. Segment Routing allows one to enforce a flow through any topological path and service chaining while maintaining per-flow state only at the ingress node to the Segment Routing domain

The Segment Routing architecture is described in ([I-D.filsfils-rtgwg-segment-routing]) The Segment Routing control plane is agnostic to the data plane, and hence it can be applied to both MPLS (and its many variants) and IPv6.

Seamless MPLS([I-D.ietf-mpls-seamless-mpls])describes an architecture which can be used to extend MPLS networks to integrate access and core/aggregation networks into a single MPLS domain. It provides a

highly flexible and a scalable architecture and the possibility to integrate hundreds of thousands of nodes.

This document describes the possibility of applying the segment routing technology to the IP RAN scenario. The segment routing could simplify the network complexity in case of IP RAN. LDP and RSVP-TE signaling protocols are not required, and the end-to-end service deployment can be achieved very easily.

2. Conventions and Abbreviations

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].

The following notations and abbreviations are used throughout this draft.

- o ASG: Aggregation Site/Service Gateway
- o BS: Base Station
- o CSG: Cell Site Gateway
- o FRR: Fast Re-Routing
- o IP RAN: Internet Protocol RAN
- o LTE: Long Term Evolution
- o RAN: Radio Access Network
- o RNC: Radio Network Controller
- o RSG: Radio Service Gateway
- o SR: Segment Routing

3. IP RAN Network Architecture

3.1. IP RAN Network Scenario

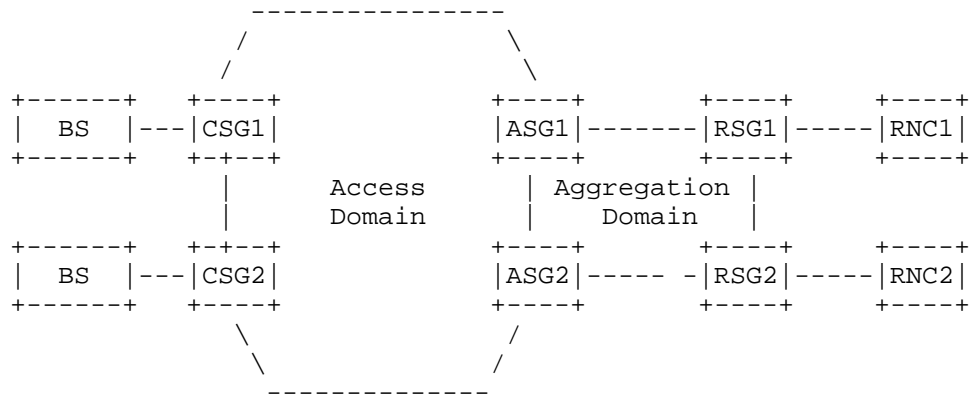


Figure 1: IP RAN Network Scenario

A typical mobile backhaul network is shown as figure 1. In the mobile backhaul network, being different from the typical access devices(DSLAM, MSAN), CSGs and RSGs of the mobile backhaul network needs to support rich MPLS features such as path design, protection switch, OAM, etc.

3.2. Requirements for IP RAN network

- (1) End-to-end transport LSP: MPLS based forwarding SHALL be provided by the Seamless MPLS based infrastructure between any nodes. The MPLS based service could be setup by L3VPN, L2VPN or pseudo wire.
- (2) OAM: The Seamless MPLS architecture should propose unified OAM mechanisms to satisfy the requirements of the end-to-end services.
- (3) Protection: The protection switch mechanism has been provided in IP RAN network to achieve convergence in 50 ms.
- (4) Scalability: With the proliferation of 3G/LTE, more and more node-Bs are deployed. IP/MPLS equipment in IP RAN network are very huge. In addition, there is more complex configuration for IP RAN network, because of the richer MPLS TE properties/features requirements. So there is more challenge in scaling the IP RAN network.
- (5) Security: The session security should be better or at least as good as in traditional IP/MPLS network.

- (6) **Survivability:** The survivability should be better or at least as good as in traditional IP/MPLS network.
- (7) **Flexibility and Overheads:** The additional overheads, if any, due to using SR should be offset by the flexibility provided by the SR in IP RANs.

4. Benefit for segment routing in IP RAN network

- (1) **Simplify end-to-end LSP tunnel establishment:** The data plane in IP RAN network is MPLS based forwarding. Segment routing technology is based on MPLS data plane, and there is no change for MPLS forwarding, so the data plane in IP RAN could use the segment routing forwarding technology. Segment routing simplify the control plane by IGP protocol distribution, there is no need for RSVP-TE and LDP signaling protocol. RSVP-TE, LDP protocol usually run in an AS, while IP-RAN network may cross AS domains. Therefore the cross-AS issue should be considered in the IP-RAN, and this is a very complex issue. Segment routing uses IGP protocol to distribute SID, and hence there is no cross-AS issue for segment routing. The BGP protocol could be extended to distribute SID in ([I-D.gredler-idr-bgp-ls-segment-routing-extension]) SR as well. The segment routing technology can simplify end-to-end LSP tunnel establishment.
- (2) **Network virtualization:** Service chaining could be introduced into SR domain. An SR header could be used to carry the set of forwarding or services that need to be applied to the packet. This can be achieved by creating an SR header with the desired sequence of service IDs that need to be applied to the packet.
- (3) **Unified OAM mechanism:** OAM mechanism could be implemented across AS by IGP and BGP extension of SID flooding. This is an easy-to-implement the cross-AS OAM mechanism. If the control plane is one or several centralized controller, the OAM policy can be determined by the controller, and the related OAM policy can be downloaded to the SR nodes seamlessly
- (4) **Traffic engineering:** Traffic Engineering has been widely addressed by using the IGP protocol extensions (for resources information propagation) and RSVP-TE for signaling explicit paths. Different contexts and modes have been defined (single vs. multiple domains, with or without bandwidth admission control, centralized vs. distributed path computation, etc), segment routing can help to implement traffic engineering in IP RAN network.

- (5) FRR: FRR technologies have been deployed by network operators in order to cope with link or node failures through pre-computation of backup paths. Segment routing can use the IP FRR technology to simplify MPLS-TE
FRR([I-D.francois-spring-resiliency-use-case]).
- (6) Flexible policy deployment: A key goal for SR is to steer a packet to follow a specific path through the network. It is possible to control the service performed at each SR node that is forwarding the packets. Forwarding is one such service provided by an SR node. The service policy can be applied to the packets in each SR node.
- (7) Simplification of management and operations: The complex RSVP-TE and LDP signaling protocol are not required in the IP RAN network anymore. Therefore, the configuration and operation management become much simple than tradition RSVP-TE based IP RAN network.
- (8) Centralization controller or distribution protocol: the control plane in IP RAN network can be IGP/BGP distribution protocol or centralization controller.

5. Unified Service Deployment

The centralization controller is supported in problem statement draft ([I-D.previdi-spring-problem-statement]) this section describe how centralization controller is applied to the IP RAN network for unified service deployment.

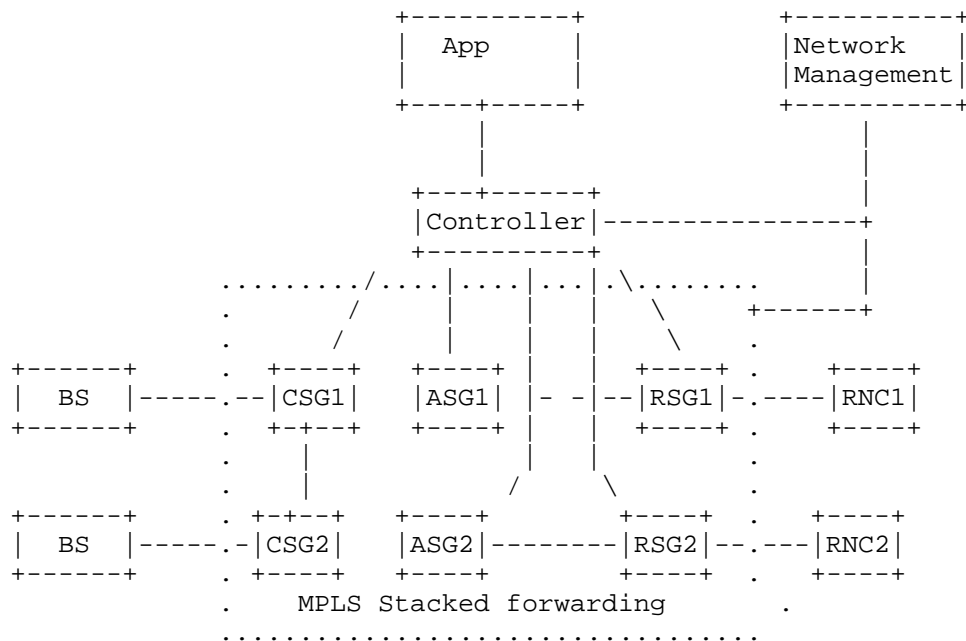


Figure 2: Centralization of Controller

Figure 2 shows an architecture for centralization of controller. The control plane is separated from the forwarding plane. IP RAN Controller is a software system that can be deployed either in a network device or a separate computer server. IP RAN controllers control the entire IP RAN network domain, the size of the domain can be defined by Network Operator based on the actual network planning and use cases. IP RAN controllers manage the IP RAN network based on the network topology, actual states and status, which are operated by the network administrator. The controller provide the northbound interface to network management system used for service deployment, monitoring, troubleshooting, fault location, etc.

CSG, ASG and RSG (we call them forwarding nodes) are only responsible for MPLS stack forwarding. RSVP-TE and LDP signaling protocol are not required in these forwarding nodes. They only need to support topology collecting and report them to controller. Forwarding nodes keep the basic routing functions in order to establish control and management channel between IP RAN Controller/NMG and all the forwarding nodes accepts network resources and states from the controller.

5.1. Requirement for Control Node

The logical centralization controller is introduced in the IP RAN network. Centralization controller is responsible for network topology collecting and label distribution based on the service.

Requirement for control node:

- (1) Control node should support collecting network topology, and managing network resource, route computing, and MPLS label distribution.
- (2) Control node support service chaining.
- (3) Control node support secure channel, and it should establish the secure connection between forwarding node.

5.2. Requirement for Forwarding Node

5.2.1. Forwarding Node Structure

The forwarding node structure in segment routing based IP RAN network is as following:

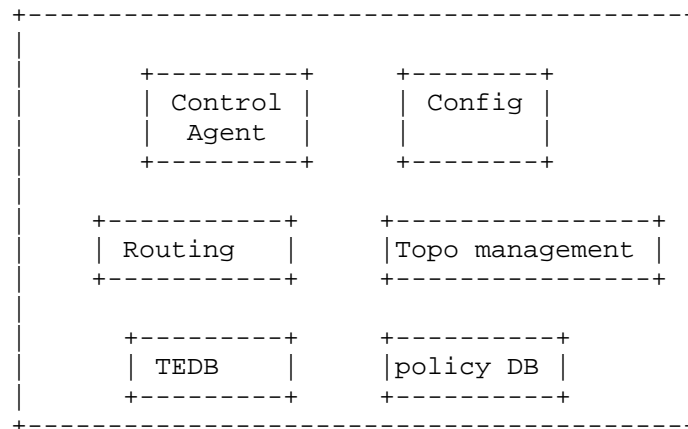


Figure 3: Forwarding node structure

The forwarding node simplifies the signaling related components, such as signal protocol component, signal label database, and a new component Control Agent is introduced to communicate with the centralization Controller.

Control Agent: control agent is used to communicate with the centralization controller. The forwarding node reports its topology and resource information, and receives the label distributed and policy through control agent. The control agent establishes the secure channel with controller. The BGP-LS protocol is recommended to use as the communication protocol between control agent and controller in this document.

Config: the config component is used for management and configuration. It is the interface with network management.

Routing: is the traditional component, is used for route computing. The routing protocol (ISIS, OSPF, and BGP) is required for the forwarding node.

Topo management: Topology management is responsible for topology computing, and topology status reporting.

TEDB: The label database.

Policy DB: policy database.

6. Security Considerations

TBD.

7. Acknowledgements

In progress.

8. IANA Considerations

This is no IANA request for this document.

9. Normative References

[I-D.filsfils-rtgwg-segment-routing]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.

[I-D.francois-spring-resiliency-use-case]

Francois, P., Filsfils, C., Decraene, B., and R. Shakir, "Use-cases for Resiliency in SPRING", draft-francois-spring-resiliency-use-case-02 (work in progress), April 2014.

[I-D.gredler-idr-bgp-ls-segment-routing-extension]

Gredler, H., Ray, S., Previdi, S., Filsfils, C., Chen, M., and J. Tantsura, "BGP Link-State extensions for Segment Routing", draft-gredler-idr-bgp-ls-segment-routing-extension-01 (work in progress), February 2014.

[I-D.ietf-mpls-seamless-mpls]

Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[I-D.previdi-spring-problem-statement]

Previdi, S., Filsfils, C., Decraene, B., Litkowski, S., Horneffer, M., Geib, R., Shakir, R., and R. Raszuk, "SPRING Problem Statement and Requirements", draft-previdi-spring-problem-statement-04 (work in progress), April 2014.

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

Authors' Addresses

Bhumip Khasnabish
ZTE TX Inc.
55 Madison Avenue, Suite 160
Morristown, New Jersey 07960
USA

Phone: +001-781-752-8003
Email: bhumip.khasnabish@ztetx.com, vumipl@gmail.com

Fangwei Hu
ZTE Corporation
No.889 Bibo Rd
Shanghai 201203
China

Phone: +86 21 68897637
Email: hu.fangwei@zte.com.cn

LUIS MIGUEL CONTRERAS MURILLO
Telefonica I+D
Distrito Telefonica, Edificio Sur 3, Planta 3
Madrid 28050
Spain

Phone: +86 21 68896273
Email: lmcm@tid.es

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 2, 2015

Z. Li
Z. Zhuang
Huawei Technologies
July 1, 2014

Use Cases and Framework of Service-Oriented MPLS Path Programming (MPP)
draft-li-spring-mpls-path-programming-00

Abstract

Source Packet Routing in Networking (SPRING) architecture for unicast traffic has been proposed to cope with the use cases in traffic engineering, fast re-reroute, service chain, etc. It can leverage existing MPLS dataplane without any modification. In fact, the label stack capability in MPLS would have been utilized well to implement flexible path programming to satisfy all kinds of requirements of service bearing. But in the distributed environment, the flexible programming capability is difficult to implement and always confined to reachability. As the introducing of central control in the network, the flexible MPLS programming capability becomes possible owing to two factors: 1. It becomes easier to allocate label for more purposes than reachability; 2. It is easy to calculate the MPLS path in a global network view. Moreover, the MPLS path programming capability can be utilized to satisfy more requirements of service bearing in the service layer which is defined as service-oriented MPLS path programming. This document defines the concept of MPLS path programming, then proposes use cases, architecture and protocol extension requirements in the service layer for the Source Packet Routing in Networking (SPRING) architecture.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 2, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. Programming Capability of MPLS Path	4
3.1. History Review	4
3.2. Gap Analysis of Segment Routing	5
4. Use Cases of Service-Oriented MPLS Path Programming	6
4.1. Use Cases for Unicast Service	6
4.1.1. Basic Reachability	6
4.1.2. VPN Identification	6
4.1.3. ECMP(Equal Cost Multi-Path)	6
4.1.4. Service OAM	7
4.1.5. Traffic Steering	7
4.2. Use Cases of Multicast Service	7
4.2.1. Basic Reachability	8
4.2.2. MVPN Identification	8
4.2.3. Source Identification	8
4.3. Use Cases of MPLS Virtual Network	8
4.4. Summary	9
5. Framework of Service-Oriented MPLS Path Programming	9
5.1. Central Control for MPLS Path Programming	9
5.2. BGP-based MPLS Segment Distribution	10
5.3. MPLS Service Path Programming	11
5.3.1. Label Combination and Download of MPLS Path	11
5.3.2. Mapping of Service Path to Service Path	11

5.4. Compatibility	12
5.5. Protocol Extensions Requirements	12
5.5.1. BGP	12
5.5.2. I2RS	12
6. IANA Considerations	12
7. Security Considerations	12
8. References	12
8.1. Normative References	12
8.2. Informative References	13
Authors' Addresses	14

1. Introduction

Source Packet Routing in Networking (SPRING) architecture for unicast traffic has been proposed to cope with the use cases in traffic engineering, fast re-reroute, service chain, etc. It can leverage existing MPLS dataplane without any modification. In fact, the label stack capability in MPLS would have been utilized well to implement flexible path programming to satisfy all kinds of requirements of service bearing. But in the distributed environment, the flexible programming capability is difficult to implement and always confined to reachability. As the introducing of central control in the network, the flexible MPLS programming capability becomes possible owing to two factors: 1. It becomes easier to allocate label for more purposes than reachability; 2. It is easy to calculate the MPLS path in a global network view. Moreover, the MPLS path programming capability can be utilized to satisfy more requirements of service bearing in the service layer which is defined as service-oriented MPLS path programming. This document defines the concept of MPLS path programming, then proposes use cases, architecture and protocol extension requirements in the service layer for the Source Packet Routing in Networking (SPRING) architecture.

2. Terminology

BGP: Border Gateway Protocol

BUM: Broadcast, Unknown unicast and Multicast

EVPN: Ethernet VPN

FRR: Fast Re-Route

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MPP: MPLS Path Programming

MVPN: Multicast VPN

RR: Route Reflector

SDN: Software-Defined Network

SR-path: Segment Routing Path

3. Programming Capability of MPLS Path

MPLS path is composed by label stacks. Since in the label stack the labels in different layers can represent different meaning and the depth of the label stack can be unlimited in theory, it is possible can make up all kinks of MPLS paths based on the combination of labels. If we look on the combination of MPLS labels as programming, it is can be seen that the MPLS path has high programming capability.

3.1. History Review

The solutions based on MPLS label stack has been widely deployed. For example, in the scenario of Options C inter-AS VPN ([RFC4364]), we assume that LDP over TE is used as the transport tunnel and the TE tunnel starts at the ingress PE, following label stack can be composed by the ingress PE for MPLS path to bear VPN service:

VPN Prefix	BGP	LDP	RSVP-TE
Label	Label	Label	Label

If facility FRR ([RFC4090]) is deployed for the MPLS TE tunnel, once the failure happens, additional label will be pushed for the label stack which is shown as follows:

VPN Prefix	BGP	LDP	RSVP-TE	BYPASS FRR
Label	Label	Label	Label	Label

The combination of labels in the above label stack is not simpler than the existing segment routing solution which composes the segment routing path through combination of segments. In fact, this is also a use case of source packet routing. But the combination is not as flexible as the segment routing since the combination of labels is always to cope with the reachability issue with limited capability in the distributed environment as follows:

1. Each label in the label stack is always binded with the reachability to a specific prefix. That is, the purpose of the label binding is limited.
2. It is difficult to implement flexible path calculation based on policy or constraints. For example, MPLS TE proposes rich set of traffic engineering attributes for transport. But it needs complex configurations in each ingress node in an unscalable way. That is, the path calculation and composition capability is limited.

As more concepts on MPLS label are proposed such as entropy label, source label, segment routing, etc., the purpose of label binding expands and the combination of labels can become more flexible. MPLS path programming capability becomes more realistic to satisfy more application scenarios.

3.2. Gap Analysis of Segment Routing

Segment Routing ([I-D.filsfils-spring-segment-routing]) is a typical example of MPLS path programming. The segment based on MPLS label is to represent nodes or agencies in the network. Through the collected information of network segments and path calculation based on the service requirement in the central controller, there will be flexible segment routing paths for the usage of traffic engineering. The SR-path can be advertised to the ingress node through PCE extensions. ([I-D.sivabalan-pce-segment-routing]).

Segment routing can implement source packet routing with high flexibility. On the other hand, there are multiple layers for MPLS path to bear services which is shown in the following figure:

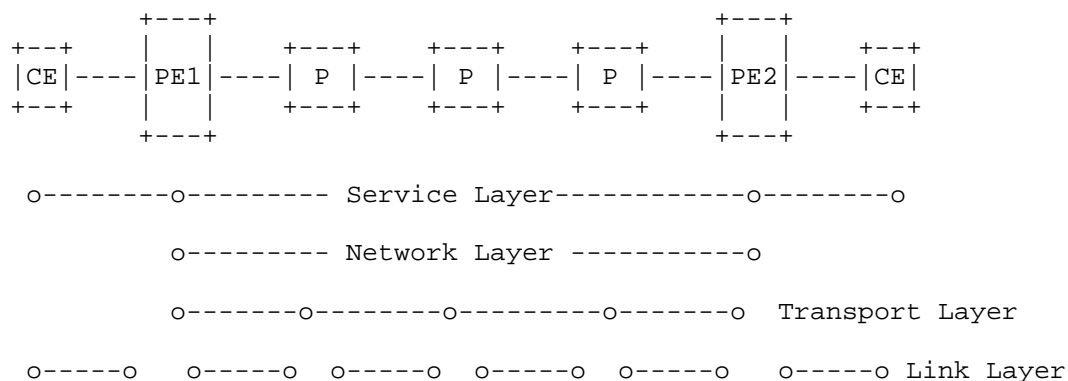


Figure 1: Multiple Layers of Service Bearing

Now the segment routing is to provide the source packet routing in the transport layer. We can call this type of source packet routing as Transport-Oriented MPLS path programming. There will be more application scenarios which needs the source packet routing in the service layer and network layer. We call these types of source packet routing as Service-Oriented MPLS path programming.

4. Use Cases of Service-Oriented MPLS Path Programming

4.1. Use Cases for Unicast Service

4.1.1. Basic Reachability

The basic reachability for VPN service is to allocate label to specific prefix including IP address or MAC address. MPLS path is as follows (using L3VPN as the example):

```
+-----+
|VPN Prefix| ---> Transport
|  Label   |      Tunnel
+-----+
```

4.1.2. VPN Identification

There are several use cases which need to indentify the VPN the packet belongs to in the forwarding plane such as the egress PE node protection for VPN ([I-D.zhang-l3vpn-label-sharing]). MPLS Path can be as follows:

```
+-----+-----+
|VPN Prefix| VPN   | ---> Transport
|  Label   | Label  |      Tunnel
+-----+-----+
```

4.1.3. ECMP(Equal Cost Multi-Path)

In order to satisfy ECMP to take full advantage of link bandwidth in the network, the entropy label ([RFC6790]) can be encapsulated. MPLS path can be as follows:

```
+-----+-----+-----+
| Entropy |VPN Prefix| VPN   | ---> Transport
|  Label  | Label   | Label  |      Tunnel
+-----+-----+-----+
```

4.1.4. Service OAM

OAM is an important requirement for the service. The performance metrics should be measured against the Service Level Agreement (SLA) for the user. Now there are relatively complete and mature OAM mechanism for the point-to-point service. But for LDP LSP, owing to the MP2P model it is difficult to identify the flow from a specific PE based on the label. Source label has been proposed as a possible solution ([I-D.chen-mpls-source-label]). When the source label is applied, MPLS path can be as follows:

-----+	-----+	-----+	-----+	-----+						
	Entropy		VPN Prefix		VPN		Source		---	Transport
	Label		Label		Label		Label			Tunnel
+	-----+	-----+	-----+	-----+						

4.1.5. Traffic Steering

Service traffic may span multiple ASes. It is an important use case to steer traffic at ASBR in an AS to specific ASBR in neighboring AS. There are possible solutions for this type of traffic steering:

1. Traffic Steering based on Transport Tunnel

This method looks on the segment between two ASBRs as the extension of the transport tunnel in an AS. It can steer the traffic through the specific path to the neighboring AS.

2. Traffic Steering in Service/Network Layer

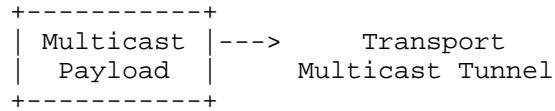
This method is to directly encapsulate the service flow with the steering label in the ingress PE before it enters into the transport tunnel. [I-D.filsfils-spring-segment-routing-central-epe] illustrates the application of Segment Routing to solve the Egress Peer Engineering (EPE) requirement. When this method is applied, the MPLS path can be as follows:

-----+												
	Entropy		Steering		VPN Prefix		VPN		Source		---	Transport
	Label		Label		Label		Label		Label			Tunnel
-----+												

4.2. Use Cases of Multicast Service

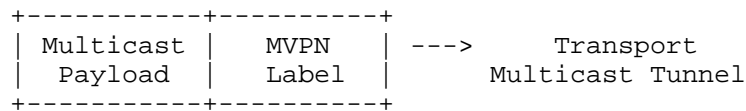
4.2.1. Basic Reachability

When MPLS multicast tunnel is applied for the multicast service in BGP-based MVPN, VPLS or EVPN, the basic MPLS path can be as follows:



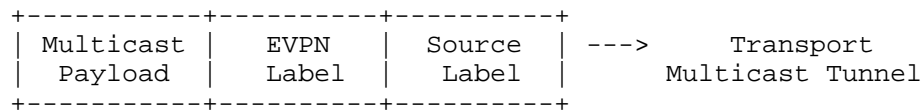
4.2.2. MVPN Identification

When multiple MVPNs shares the MPLS multicast tunnel, it is necessary to encapsulate the label to identify specific MVPN([RFC6514]). The MPLS path can be as follows:



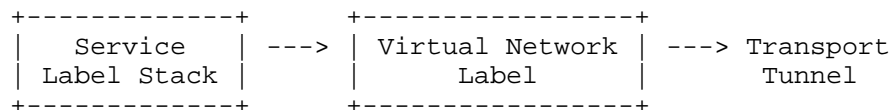
4.2.3. Source Identification

In order to implement the split horizon or C-MAC learning in the forwarding plane when MPLS multicast is to bear BUM traffic in L2VPN, it is necessary to introduce the label to identify the source of the BUM traffic([I-D.li-l2vpn-segment-evpn]). The MPLS path is as follows:



4.3. Use Cases of MPLS Virtual Network

The framework of MPLS virtual network has been proposed in [I-D.li-mpls-network-virtualization-framework]. When the unicast service or the multicast service enters into the transport tunnel, it may take different MPLS virtual network identified by the MPLS label for the purpose of QoS routing, security or virtual operations. The MPLS path is as follows:



4.4. Summary

Service-oriented MPLS path programming can make full use of flexible combination of MPLS labels to satisfy different requirements for the service flow. Based on the above proposed use cases, MPLS path can be composed adopting part or whole labels for these use cases based on the service requirement. Besides this, more flexible MPLS label combination may be provided:

1. Hierarchical process or multiple repeated process: The label for the same usage can exist in different layers. Or the process identified by the label can exist in multiple nodes along the path. Then the labels for the same usage can be encapsulated several times in the label stack. The encapsulation can be as follows (using SERVICE LABEL to identify the label for the same service process in different layers):

+-----+	+-----+	+-----+	+-----+	+-----+	+-----+
SERVICE	VPN Prefix	SERVICE	VPN	SERVICE	Tunnel
LABEL	Label	LABEL	Label	LABEL	Label
+-----+	+-----+	+-----+	+-----+	+-----+	+-----+

2. Special-purpose label indicator: Since the label in the service-oriented MPLS programming is for special-purpose process, it may need a special purpose label to indicate the usage of the label followed the special-purpose labels. For example, the ELI(Entropy Label Indicator) is introduced for the entropy label. This may introduce more labels for the combination.

This document is not to define all possible use cases for the service-oriented path programming. The new use cases can be defined in the future independent document.

5. Framework of Service-Oriented MPLS Path Programming

5.1. Central Control for MPLS Path Programming

Central control plays an important role in MPLS path programming. It can extend the MPLS path programming capability easily. There are two important functionalities for the central control:

1. Central controlled MPLS label allocation: Label can be allocated centrally for special usage other than reachability. These labels can be used to compose MPLS path. We call it as MPLS Segment.
2. Central controlled MPLS path programming: Central controller can calculate path in a global network view and implement the MPLS path

programming based on the collected information of MPLS segments to satisfy different requirements of services.

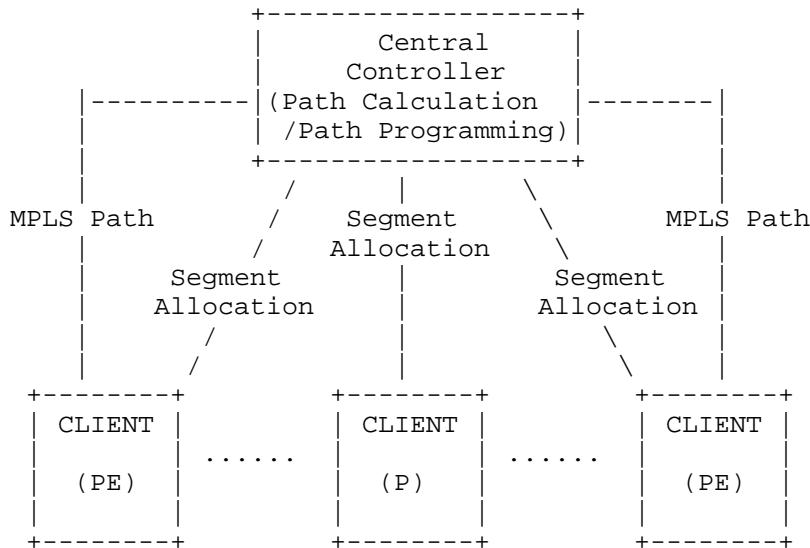


Figure 2 Central Control for MPLS Path Programming

There are two types of MPLS path: Transport-Oriented MPLS Path and Service-Oriented MPLS Path. For the transport-oriented MPLS path, segment routing is the typical solution: MPLS segment distribution is done by IGP extensions ([I-D.ietf-isis-segment-routing-extensions]) and [I-D.ietf-ospf-segment-routing-extensions]); the programmed MPLS path can be downloaded through PCEP extensions from PCE to PCC([I-D.sivabalan-pce-segment-routing]). For the service-oriented MPLS path programming, it not only includes composing the MPLS path in the service and network layer, but also includes determining the mapping of the service path to the transport path. Since the process corresponding to the label in the service label stack is always located at the PE nodes, BGP extensions can be introduced for service-oriented path programming.

5.2. BGP-based MPLS Segment Distribution

1. Label Allocation

There are two types of label used for MPLS segments:

1) Local Label: The service process is done locally. The label can be allocated by the local PE which provides the process.

2) Global Label: The service process is common in multiple PEs. This means the label has global meaning. The label allocation can be done by the central controller. The global label work can refer to [I-D.li-mpls-global-label-framework].

2. Label Mapping Distribution

BGP extensions can be used to distribution label mapping. Regarding to the above two types of label allocation, the process is as follows:

1) Local Label Mapping: BGP can directly distribute the label mapping from the local PE to peer PEs. The local PE can also only distribute the label mapping to central controller. Then the central controller re-distribute the label mapping to other PEs. In this method, the central controller plays the role of traditional RR.

2) Global Label Mapping: The label mapping for the service can be directly distributed by the central controller to multiple PEs. It can be done by BGP extensions.

5.3. MPLS Service Path Programming

5.3.1. Label Combination and Download of MPLS Path

According to the service requirements, the central controller can combine MPLS segments flexibly. Then it can download the service label combination for specific prefix related with the service. The BGP extensions can be reused to download the programmed MPLS path.

5.3.2. Mapping of Service Path to Service Path

Since the transport path is also to satisfy the service bearing the requirement, it can reuse the existing MPLS tunnel technology or it can also be programmed according to traffic engineering requirements of service. Then there needs to be implements the mapping of the service path to the transport path. There are two ways to implement the mapping:

1. BGP Extensions: Through the community attribute of BGP, the identifier of the transport path can be carrier when distribute label stack for a specific prefix.

2. I2RS Extensions: I2RS can be used to download route policy to the client node. Based on the policy, the client node can implement the required mapping.

5.4. Compatibility

When the MPLS path programming is done the central controller and downloaded through BGP extensions to the Client node, the path SHOULD have higher priority than the path calculated on the Client node's own.

5.5. Protocol Extensions Requirements

5.5.1. BGP

REQ 01: BGP extensions SHOULD be introduced to distribute local label mapping for specific process.

REQ 02: BGP extensions SHOULD be introduced to distribute global label mapping for specific process.

REQ 03: BGP extensions SHOULD be introduced to download label stack for service-oriented MPLS path.

REQ 04: BGP extensions SHOULD be introduced to carry the identifier of the transport MPLS path with service MPLS path to implement the mapping.

5.5.2. I2RS

REQ 11: I2RS clients SHOULD provide interface to I2RS agent to download policy to implement the mapping of the service path to the transport path.

6. IANA Considerations

This document makes no request of IANA.

7. Security Considerations

TBD.

8. References

8.1. Normative References

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

8.2. Informative References

- [I-D.chen-mpls-source-label]
Chen, M., Xu, X., Li, Z., Fang, L., and G. Mirsky,
"MultiProtocol Label Switching (MPLS) Source Label",
draft-chen-mpls-source-label-03 (work in progress), April
2014.
- [I-D.filsfils-spring-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B.,
Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R.,
Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe,
"Segment Routing Architecture", draft-filsfils-spring-
segment-routing-03 (work in progress), June 2014.
- [I-D.filsfils-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Patel, K., Aries, E.,
shaw@fb.com, s., Ginsburg, D., and D. Afanasiev, "Segment
Routing Centralized Egress Peer Engineering", draft-
filsfils-spring-segment-routing-central-epe-01 (work in
progress), May 2014.
- [I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H.,
Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS
Extensions for Segment Routing", draft-ietf-isis-segment-
routing-extensions-02 (work in progress), June 2014.
- [I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,
Shakir, R., Henderickx, W., and J. Tantsura, "OSPF
Extensions for Segment Routing", draft-ietf-ospf-segment-
routing-extensions-00 (work in progress), June 2014.
- [I-D.li-l2vpn-segment-evpn]
Li, Z., Yong, L., and J. Zhang, "Segment-Based
EVPN(S-EVPN)", draft-li-l2vpn-segment-evpn-01 (work in
progress), February 2014.
- [I-D.li-mpls-global-label-framework]
Li, Z., Zhao, Q., and T. Yang, "A Framework of MPLS Global
Label", draft-li-mpls-global-label-framework-01 (work in
progress), February 2014.
- [I-D.li-mpls-global-label-usecases]
Li, Z., Zhao, Q., and T. Yang, "Useases of MPLS Global
Label", draft-li-mpls-global-label-usecases-01 (work in
progress), February 2014.

- [I-D.li-mpls-network-virtualization-framework]
Li, Z. and M. Li, "Framework of Network Virtualization Based on MPLS Global Label", draft-li-mpls-network-virtualization-framework-00 (work in progress), October 2013.
- [I-D.sivabalan-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and R. Raszuk, "PCEP Extensions for Segment Routing", draft-sivabalan-pce-segment-routing-02 (work in progress), October 2013.
- [I-D.zhang-l3vpn-label-sharing]
Zhang, M., Zhou, P., and R. White, "Label Sharing for Fast PE Protection", draft-zhang-l3vpn-label-sharing-02 (work in progress), June 2014.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

G. Mirsky
J. Tantsura
Ericsson
I. Varlashkin
EasyNet
June 30, 2014

Bidirectional Forwarding Detection (BFD) Directed Return Path
draft-mirsky-mpls-bfd-directed-00

Abstract

Bidirectional Forwarding Detection (BFD) is expected to monitor bi-directional paths. When forward direction of a BFD session is to monitor explicitly routed path there is a need to be able to direct far-end BFD peer to use specific path as reverse direction of the BFD session.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 1, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	3
1.1.1. Terminology	3
1.1.2. Requirements Language	3
2. Problem Statement	3
3. Direct Reverse BFD Path	3
3.1. Case of MPLS Data Plane	4
3.1.1. BFD Reverse Path TLV	4
3.1.2. Segment Routing Tunnel sub-TLV	4
3.2. Case of IPv6 Data Plane	5
4. IANA Considerations	6
4.1. TLV	6
4.2. Sub-TLV	6
5. Security Considerations	7
6. Acknowledgements	7
7. Normative References	7
Authors' Addresses	8

1. Introduction

The [RFC5880], [RFC5881], and the [RFC5883] established BFD protocol for IP networks and the [RFC5884] set rules of using BFD Asynchronous mode over IP/MPLS LSPs. All standards implicitly assume that the far-end BFD peer will use the best route regardless of route being used to send BFD control packets towards it. As result, if the near-end BFD peer sends its BFD control packets over explicit path that is diverging from the best route, then reverse direction of the BFD session is likely not to be on co-routed bi-directional path with the forward direction of the BFD session. And because BFD control packets are not guaranteed to cross the same links and nodes in both directions detection of Loss of Continuity (LoC) defect in forward direction is not guaranteed or free of positive negatives.

This document proposes to use BFD Return Path TLV extension to LSP Ping [RFC4379] to instruct the far-end BFD peer to use explicit path for its BFD control packets associated with the particular BFD session. As a special case, forward and reverse directions of the BFD session can form bi-directional co-routed associated channel.

1.1. Conventions used in this document

1.1.1. Terminology

BFD: Bidirectional Forwarding Detection

MPLS: Multiprotocol Label Switching

LSP: Label Switching Path

LoC: Loss of Continuity

1.1.2. 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 [RFC2119].

2. Problem Statement

BFD is best suited to monitor bi-directional co-routed paths. In most cases, in IP and IP/MPLS networks the best route between two IP nodes is likely to be co-routed in the stable network environment so that implicit BFD requirement is being fulfilled. If BFD is tasked to monitor unidirectional explicitly routed path, e.g. MPLS LSP, its control packets in forward direction would be in-band due to mechanism defined in [RFC5884] and [RFC5586]. But the reverse direction of the BFD session would still follow the best route and that presents following problems in regard to detecting defects on the unidirectional explicit path:

- failure detection on the reverse path cannot be interpreted as bi-directional failure and thus trigger, for example, protection switchover of the forward direction;
- if reverse direction is in Down state, the head-end node would not receive indication of forward direction failure from its far-end peer.

To address these challenges the far-end BFD peer should be instructed to use specific path for its control packets.

3. Direct Reverse BFD Path

3.1. Case of MPLS Data Plane

LSP ping, defined in [RFC4379], uses BFD Discriminator TLV [RFC5884] to bootstrap a BFD session over an MPLS LSP. This document defines a new TLV, BFD Reverse Path TLV, that must contain a single sub-TLV that can be used to carry information about reverse path for the specified in BFD Discriminator TLV session.

3.1.1. BFD Reverse Path TLV

The BFD Reverse Path TLV is an optional TLV within the LSP ping protocol. However, if used the BFD Discriminator TLV MUST be included in an Echo Request message as well. If the BFD Discriminator TLV is not present when the BFD Reverse Path TLV is included, then it MUST be treated as malformed Echo Request, as described in [RFC4379].

The BFD Reverse Path TLV carries the specified path that BFD control packets of the BFD session referenced in the BFD Discriminator TLV are required to follow. The format of the BFD Reverse Path TLV is as presented in Figure 1.

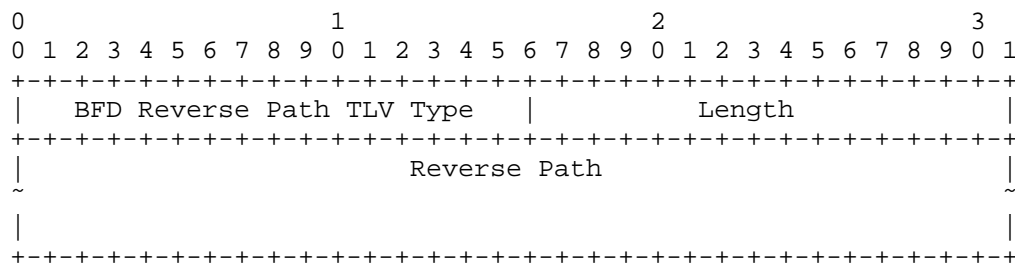


Figure 1: BFD Reverse Path TLV

BFD Reverse Path TLV Type is 2 octets in length and value to be assigned by IANA.

Length is 2 octets in length and defines the length in octets of the Reverse Path field.

3.1.2. Segment Routing Tunnel sub-TLV

With MPLS data plane explicit path can be either Static or RSVP-TE LSP, or Segment Routing tunnel. In case of Static or RSVP-TE LSP [RFC7110] defined sub-TLVs to identify explicit return path. For the Segment Routing with MPLS data plane case a new sub-TLV is defined in this document as presented in Figure 2.

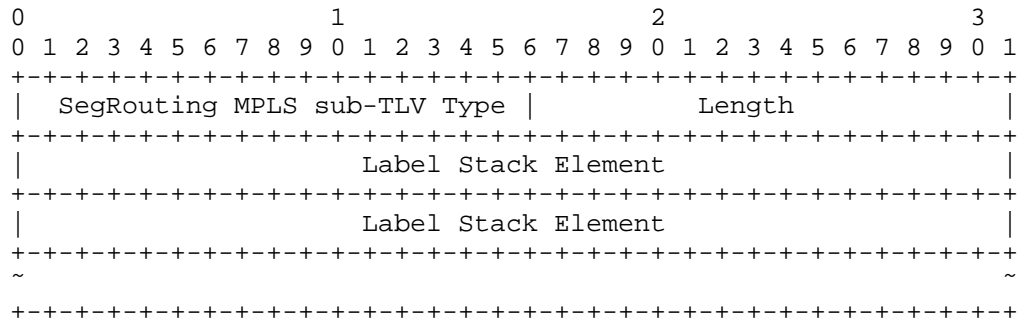


Figure 2: Segment Routing MPLS Tunnel sub-TLV

The Segment Routing Tunnel sub-TLV Type is two octets in length, and will be allocated by IANA.

The Segment Routing Tunnel sub-TLV MAY be used in Reply Path TLV defined in [RFC7110]

3.2. Case of IPv6 Data Plane

IPv6 can be data plane of choice for Segment Routed tunnels [I-D.previdi-6man-segment-routing-header]. In such networks the BFD Reverse Path TLV described in Section 3.1.1 can be used as well. IP networks, unlike IP/MPLS, do not require use of LSP ping with BFD Discriminator TLV[RFC4379] to bootstrap BFD session. But to specify reverse path of a BFD session in IPv6 environment the BFD Discriminator TLV MUST be used along with the BFD Reverse Path TLV. The BFD Reverse Path TLV in IPv6 network MUST include sub-TLV.

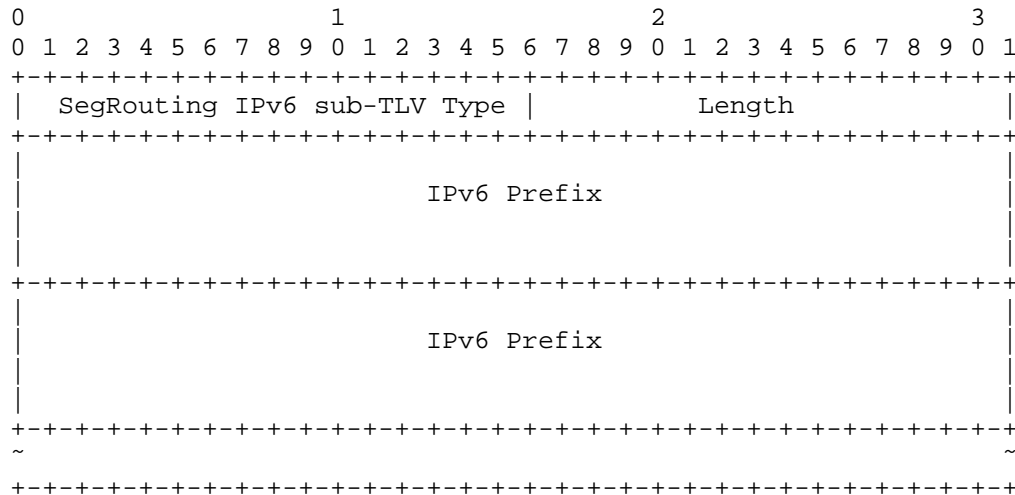


Figure 3: Segment Routing IPv6 Tunnel sub-TLV

4. IANA Considerations

4.1. TLV

The IANA is requested to assign a new value for BFD Reverse Path TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

Value	Description	Reference
X (TBD1)	BFD Reverse Path TLV	This document

Table 1: New BFD Reverse Type TLV

4.2. Sub-TLV

The IANA is requested to assign one new sub-TLV type from "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "Sub-TLVs for TLV Type 1" sub-registry.

Value	Description	Reference
X (TBD2)	Segment Routing MPLS Tunnel sub-TLV	This document
X (TBD3)	Segment Routing IPv6 Tunnel sub-TLV	This document

Table 2: New Segment Routing Tunnel sub-TLV

5. Security Considerations

Security considerations discussed in [RFC5880], [RFC5884], and [RFC4379], apply to this document.

6. Acknowledgements

7. Normative References

- [I-D.previdi-6man-segment-routing-header]
Previdi, S., Filsfils, C., Field, B., and I. Leung, "IPv6 Segment Routing Header (SRH)", draft-previdi-6man-segment-routing-header-01 (work in progress), June 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.
- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, June 2010.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

[RFC7110] Chen, M., Cao, W., Ning, S., Jounay, F., and S. Delord,
"Return Path Specified Label Switched Path (LSP) Ping",
RFC 7110, January 2014.

Authors' Addresses

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

Jeff Tantsura
Ericsson

Email: jeff.tantsura@ericsson.com

Ilya Varlashkin
EasyNet

Email: Ilya.Varlashkin@easynet.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 31, 2014

X. Xu
Z. Li
Huawei
H. Shah
Ciena
L. Contreras
Telefonica I+D
June 29, 2014

Service Function Chaining Use Case for SPRING
draft-xu-spring-sfc-use-case-02

Abstract

This document describes a particular use case for SPRING where the Segment Routing mechanism is leveraged to realize the service path layer functionality of the Service Function Chaining (i.e, steering traffic through the service function path).

Status of This Memo

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

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

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

This Internet-Draft will expire on December 31, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

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

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Requirements Language	3
2. Terminology	3
3. SFC Use Case	3
3.1. SFC in MPLS-SR Case	3
3.2. SFC in IPv6-SR Case	4
4. Acknowledgements	5
5. IANA Considerations	5
6. Security Considerations	5
7. References	5
7.1. Normative References	5
7.2. Informative References	6
Authors' Addresses	6

1. Introduction

When applying a particular Service Function Chaining (SFC) [I-D.quinn-sfc-arch] to the traffic selected by the service classifier, the traffic need to be steered through an ordered set of service nodes in the network. This ordered set of service nodes indicates the service function path which is actually the instantiation of the above SFC in the network. Furthermore, additional information about the traffic (a.k.a. metadata) which is helpful for enabling value-added services may need to be carried across those service nodes within the SFC instantiation. As mentioned in [I-D.rijsman-sfc-metadata-considerations] "...it is important to make a distinction between fields which are used at the service path layer to identify the Service Path Segment, and additional fields which carry metadata which is imposed and interpreted at the service function layer. Combining both types of fields into a single header should probably be avoided from a layering point of view. "

Segment Routing (SR) [I-D.filsfils-spring-segment-routing] is a source routing paradigm which can be used to steer traffic through an ordered set of routers. SR can be applied to the MPLS data plane [I-D.gredler-spring-mpls] and the IPv6 data plane [I-D.filsfils-spring-segment-routing-mpls] and the IPv6 data plane [I-D.previdi-6man-segment-routing-header].

This document describes a particular use case for SPRING where the SR mechanism is leveraged to realize the service path layer

functionality of the SFC (i.e, steering traffic through the service function path).

1.1. Requirements Language

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

2. Terminology

This memo makes use of the terms defined in [I-D.filsfils-spring-segment-routing] and [I-D.quinn-sfc-arch].

3. SFC Use Case

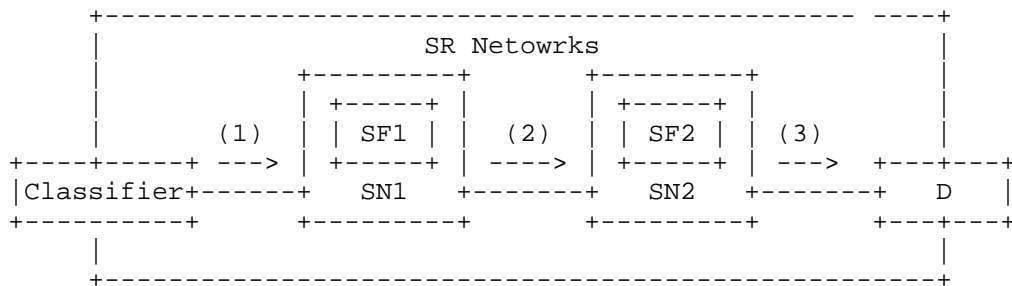


Figure 1: Service Function Chaining in SR Networks

As shown in Figure 1, assume SN1 and SN2 are two SR-capable nodes meanwhile they are service nodes which offer service function SF1 and SF2 respectively. In addition, they have allocated and advertised segment IDs (SID) for the service functions they are offering. For example, SN1 allocates and advertises an SID, i.e., SID(SF1) for service function SF1 while SN2 allocates and advertises an SID, i.e., SID(SF2) for service function SF2. These SIDs which are used to indicate service functions are referred to as Service Function SIDs. In addition, assume the node SIDs for SN1 and SN2 are SID(SN1) and SID(SN2) respectively.

How to steer a packet through a service function path in both MPLS-SR and IPv6-SR cases is illustrated in the following two sub-sections respectively.

3.1. SFC in MPLS-SR Case

In the MPLS-SR case, those service function SIDs as mentioned above would be interpreted as local MPLS labels. Meanwhile, to simplify

the illustration in this document, those node SIDs as mentioned above would be interpreted as MPLS global labels.

Now assume a given packet destined for destination D is required to go through a service function chain {SF1, SF2} before reaching its final destination D. The service classifier therefore would attach a segment list {SID(SN1), SID(SF1), SID(SN2), SID(SF2)} to the packet. This segment list is actually represented by a MPLS label stack. In addition, the service classifier could optionally impose metadata on the packet through the Network Service Header (NSH) [I-D.quinn-sfc-nsh]. Here the Service Path field within the NSH would not be used for the path selection purpose anymore and therefore it MUST be set to a particular value to indicate such particular usage. In addition, the service index value within the NSH is set to 2 since there are two service nodes within the service function path. How to impose the NSH on a MPLS packet is outside the scope of this document. When the encapsulated packet arrives at SN1, SN1 would know which service function should be performed according to SID (SF1). If a NSH is carried in that packet, SN1 could further consume the metadata contained in the NSH and meanwhile decrease the service index value within the NSH by one. When the encapsulated packet arrives at SN2, SN2 would do the similar action as what has been done by SN1. Furthermore, since SN2 is the last service node within the service function path, SN2 MUST strip the NSH (if it has been imposed) before sending the packet to D.

3.2. SFC in IPv6-SR Case

In the IPv6-SR case, those service function SIDs as mentioned above would be interpreted as IPv6 link-local addresses while those node SIDs as mentioned above would be interpreted as IPv6 global unicast addresses.

Now assume a given IPv6 packet destined for destination D is required to go through a service function chain {SF1, SF2} before reaching its final destination D. The service classifier therefore would attach a SR header containing a segment list {SID(SF1), SID(SN2), SID(SF2), SID(D)} to the IPv6 packet. This segment list is actually represented by an ordered list of IPv6 addresses. The IPv6 destination address is filled with SID(SN1). In addition, the service classifier could optionally impose metadata on the above IPv6 packet through the NSH and meanwhile carry the original IPv6 source address in the Original Source Address field of the packet. When the above IPv6 packet arrives at SN1, SN1 would know which service function should be performed according to SID (SF1). If a NSH is carried in that packet, SN1 could further consume the metadata contained in the NSH and meanwhile decrease the service index value within the NSH by one. When the packet arrives at SN2, SN2 would do

the similar action as what has been done by SN1. Furthermore, since SN2 is the second last node in the segment list, SN2 should strip the SR header and meanwhile fill in the IPv6 source address with the Original Source Address (if available) before sending the packet towards D. Besides, since SN2 is the last service node within the service path, SN2 MUST strip the NSH (if it has been imposed) before sending the packet to D.

4. Acknowledgements

TBD.

5. IANA Considerations

TBD.

6. Security Considerations

This document does not introduce any new security risk.

7. References

7.1. Normative References

[I-D.filsfils-spring-segment-routing-mpls]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-filsfils-spring-segment-routing-mpls-02 (work in progress), June 2014.

[I-D.gredler-spring-mpls]

Gredler, H., Rekhter, Y., Jalil, L., Kini, S., and X. Xu, "Supporting Source/Explicitly Routed Tunnels via Stacked LSPs", draft-gredler-spring-mpls-06 (work in progress), May 2014.

[I-D.previdi-6man-segment-routing-header]

Previdi, S., Filsfils, C., Field, B., and I. Leung, "IPv6 Segment Routing Header (SRH)", draft-previdi-6man-segment-routing-header-01 (work in progress), June 2014.

[I-D.quinn-sfc-arch]

Quinn, P. and J. Halpern, "Service Function Chaining (SFC) Architecture", draft-quinn-sfc-arch-05 (work in progress), May 2014.

[I-D.quinn-sfc-nsh]

Quinn, P., Guichard, J., Fernando, R., Surendra, S., Smith, M., Yadav, N., Agarwal, P., Manur, R., Chauhan, A., Elzur, U., McConnell, B., and C. Wright, "Network Service Header", draft-quinn-sfc-nsh-02 (work in progress), February 2014.

[I-D.rijsman-sfc-metadata-considerations]

Rijsman, B. and J. Moisand, "Metadata Considerations", draft-rijsman-sfc-metadata-considerations-00 (work in progress), February 2014.

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

7.2. Informative References

[I-D.filsfils-spring-segment-routing]

Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-spring-segment-routing-03 (work in progress), June 2014.

Authors' Addresses

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Zhenbin Li
Huawei

Email: lizhenbin@huawei.com

Himanshu Shah
Ciena

Email: hshah@ciena.com

Luis M. Contreras
Telefonica I+D
Ronda de la Comunicacion, s/n
Sur-3 building, 3rd floor
Madrid, 28050
Spain

Email: luismiguel.contrerasmurillo@telefonica.com
URI: <http://people.tid.es/LuisM.Contreras/>