Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 30, 2013

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 28, 2013

Segment Routing Use Cases
draft-filsfils-rtgwg-segment-routing-use-cases-00

Abstract

Segment Routing (SR) leverages the source routing and tunneling paradigms. 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. 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.

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. Segment Routing can also be applied to IPv6 with a new type of routing extension header.

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 30, 2013.

Copyright Notice

Table of Contents

1. Introduction

In this document, we document various SR use-cases.

Section 2 illustrates the ability to tunnel traffic towards remote service points without any other protocol than the IGP.

Section 3 reports various FRR use-cases leveraging the SR functionality.

Section 4 and Section 5 document traffic-engineering use-cases, respectively without and with a notion of bandwidth admission control.

Section 6 documents the co-existence and interworking with MPLS Signaling protocols.

Section 7 illustrates OAM use-cases.

The objective of this document is to illustrate the properties and benefits of the SR architecture. To avoid any risk of partiality and lengthy debate, we do not include any comparison with other architectures.

1.1. Companion Documents

The main reference for this document is the SR architecture defined in [draft-filsfils-rtgwg-segment-routing-00].

IS-IS protocol extensions for Segment Routing are described in [draft-previdi-isis-segment-routing-extensions-00].

OSPF protocol extensions for Segment Routing are defined in [draft-psenak-ospf-segment-routing-extensions-00].

The PCEP protocol extensions for Segment Routing are defined in [draft-msiva-pce-pcep-segment-routing-extensions-00].

In the future, we will submit the FRR section as an independent document.

2. IGP-based MPLS Tunneling

SR, applied to the MPLS dataplane, 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. LDP and RSVP-TE signaling protocols are not required.

Note that Section 6 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.

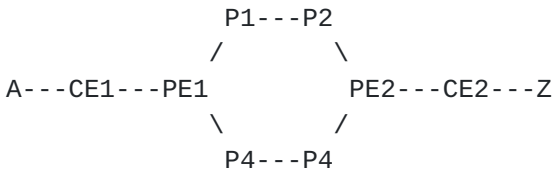


Figure 1: IGP-based MPLS Tunneling

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. FRR

3.1. Protecting a resource along the path of a Node Segment

SR leverages the technologies stemming from the IPFRR framework to provide fast recovery of end-to-end connectivity upon failures. This section assumes familiarity with Remote-LFA concepts described in [I-D.ietf-rtgwg-remote-lfa].

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

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

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

If the P and Q sets do not intersect (i.e. there is no RLFA PQ node), we propose to use a Directed LFA (DLFA) repair tunnel from S to a Q node that is adjacent to the P space ([I-D.ietf-rtgwg-remote-lfa]). The LFA repair tunnel only requires two segments: a node segment to a P node which is adjacent to the Q node and an adjacency segment from the P node to its adjacent Q node.

Thanks to the DLFA extension, we thus have a guaranteed LFA-based FRR technique for any network with symmetric IGP metrics. Future versions of the document will describe the solutions leveraging SR capabilities to provided guaranteed FRR applicability in any IGP topology.

Resolving FRR with SR has the following benefits:

- Preservation of the simplicity properties of LFA FRR ([RFC6571]).

- Preservation of the capacity planning properties (unlike SDH and other FRR solutions, the repaired packet does not go back to the next-hop or next-next-hop but uses shortest-path forwarding from a much closer release point, [RFC6571]).

Simplification of the RLFA operation: no dynamically-established directed LDP sessions to the repair nodes.

No requirement for any extra computation on top of the one required for RLFA.

Guaranteed coverage for symmetric networks,

The repair tunnel in symmetric network can be encoded efficiently with only two segments.

3.2. Protecting an adjacency segment

More details will be provided in a future version.

3.3. Protecting a node segment upon the failure of its advertising node

Referring to the below figure, let us assume:

A is identified by IP address 192.0.2.1/32 to which Node-SID 101 is attached.

B is identified by IP address 192.0.2.2/32 to which Node-SID 102 is attached

A and B host the same set of services.

Each service is identified by a local segment at each node: i.e. node A allocates a local service segment 9001 to identify a specific service S while the same service is identified by a local service segment 9002 at B. Specifically, for the sake of this illustration, let us assume that service S is a BGP-VPN service where A announces a VPN route V with BGP nhop 192.0.2.1/32 and local VPN label 9001 and B announces the same VPN route V with BGP nhop 192.0.2.2/32 and local VPN label 9002.

A generic mesh interconnects the three nodes M, Q and B.

N prefers to use the service S offered by A and hence sends its S-destined traffic with segment list {101, 9001}.

Q is a node connected to A.

Q has a method to detect the loss of node A within a few 10's of msec.

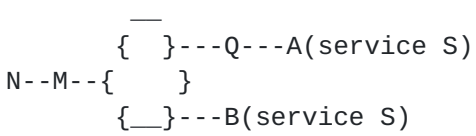


Figure 2: Service Mirroring

In that context, we would like to protect the traffic destined to service S upon the failure of node A.

The solution is built upon several components:

1. B advertises its mirroring capability for mirrored Node-SID 101
2. B pre-installs a mirroring table in order to process the packets originally destined to 101.
3. Q and any neighbor of A pre-install the Mirror_FRR LFA extension
4. All nodes implements a modified SRDB convergence upon Node-SID 101 deletion

3.3.1. Advertisement of the Mirroring Capability

B advertises a MIRROR sub-TLV in its IGP Link-State Router Capability TLV with the values (TTT=000, MIRRORED_OBJECT=101, CONTEXT_SEGMENT=10002),[draft-filsfils-rtgwg-segment-routing-00], [draft-previdi-isis-segment-routing-extensions-00] and [draft-psenak-ospf-segment-routing-extensions-00] for more details in the encodings.

Doing so, B advertises within the routing domain that it is willing to backup any traffic originally sent to Node-SID 101 provided that this rerouted traffic gets to B with the context segment 10002 directly preceding any local service segment advertised by A. 10002 is a local context segment allocated by B to identify traffic that was originally meant for A. This allows B to match the subsequent service segment (e.g. 9001) correctly.

3.3.2. Mirroring Table

We assume that B is able to discover all the local service segments allocated by A (e.g. BGP route reflection and add-path). B maps all the services advertised by A to its similar service representations. For example, service 9001 advertised by A is mapped to service 9002 advertised by B as both relate to the same service S (the same VPN route V). For example, B applies the same service treatment to a packet received with top segments {102, 10002, 9001} or with top segments {102, 9002}. Basically, B treats {10002, 9001} as a synonym of {9002}.

3.3.3. LFA FRR at the Point of Local Repair

In advance of any failure of A, Q (and any other node connected to A) learns the identity of the IGP Mirroring node for each Node-SID advertised by A (MIRROR_TLV advertised by B) and pre-installs the following new MIRROR_FRR entry:
- Trigger condition: the loss of nhop A
- Incoming active segment: 101 (a Node-SID advertised by A)
- Primary Segment processing: pop 101
  - Backup Segment processing: pop 101, push {102, 10002}
- Primary nhop: A
  - Backup nhop: primary path to node B

Upon detecting the loss of node A, Q intercepts any traffic destined to Node-SID 101, pops the segment to A (101) and push a repair tunnel {102, 10002}. Node-SID 102 steers the repaired traffic to B while context segment 10002 allows B to process the following service segment {9001} in the right context table.

3.3.4. Modified IGP Convergence upon Node deletion

Upon the failure of A, all the neighbors of A will flood the loss of their adjacency to A and eventually every node within the IGP domain will delete 192.0.2.1/32 from their RIB.

The RIB deletion of 192.0.2.1/32 at N is beneficial as it triggers the BGP FRR Protection onto the precomputed backup next-hop [draft-rtgwg-bgp-pic-01.txt].

The RIB deletion at node M, if it occurs before the RIB deletion at N, would be disastrous as it would lead to the loss of the traffic from N to A before Q is able to apply the Mirroring protection.

The solution consists in delaying the deletion of the SRDB entry for 101 by 2 seconds while still deleting the IP RIB 192.0.2.1/32 entry immediately.

The RIB deletion triggers the BGP FRR and BGP Convergence. This is beneficial and must occur without delay.

The deletion of the SRDB entry to Node-SID101 is delayed to ensure that the traffic still in transit towards Node-SID 101 is not dropped.

The delay timer should be long enough to ensure that either the BGP FRR or the BGP Convergence has taken place at N.

3.3.5. Conclusions

In our reference figure, N sends its packets towards A with the segment list {101, 9001}. The shortest-path from S to A transits via M and Q.

Within a few msec of the loss of A, Q activates its pre-installed Mirror_FRR entry and reroutes the traffic to B with the following segment list {102, 10002, 9001}.

Within a few 100's of msec, any IGP node deletes its RIB entry to A but keeps its SRDB entry to Node-SID 101 for an extra 2 seconds.

Upon deleting its RIB entry to 192.0.2.1/32, N activates its BGP FRR entry and reroutes its S destined traffic towards B with segment list {102, 9002}.

By the time any IGP node deletes the SRDB entry to Node-SID 101, N no longer sends any traffic with Node-SID 101.

The deletion of the SRDB entry to Node-SID101 is delayed to ensure that the traffic still in transit towards Node-SID 101 is not dropped.

In conclusion, the traffic loss only depends on the ability of Q to detect the node failure of its adjacent node A.

4. Traffic Engineering without Bandwidth Admission Control

This section describes traffic-engineering use-cases which do not require bandwidth admission control.

The first sub-section illustrates the use of anycast segments to express macro policies. Two examples are provided: one involving a disjointness enforcement within a so-called dual-plane network, and the other involving CoS-based policies.

The second sub-section illustrate how a head-end router can combine a distributed CSPF computation with SR. Various examples are provided where the CSPF constraint or objective is either a TE affinity, an SRLG or a latency metric.

The third sub-section illustrates how SR can help traffic-engineer outbound traffic among different external peers, overriding the best installed IP path at the egress border routers.

The fourth sub-section describes how SR can be used to express

deterministic non-ECMP path. Several techniques to compress the related segment lists are also introduced.

The fifth sub-section describes a use-case where a node attaches an Adj-SID to a set of its interfaces however not sharing the same neighbor. The illustrated benefit relates to loadbalancing.

4.1. Anycast Node Segment

The SR architecture defines an anycast segment as a segment attached to an anycast IP prefix ([RFC4786]).

The anycast node segment is an interesting tool for traffic engineering:

- Macro-policy support: anycast segments allow to express policies such as "go via plane1 of a dual-plane network" (Section 4.1.1) or "go via Region3" (Section 4.3).

- Implicit node resiliency: the traffic-engineering policy is not anchored to a specific node whose failure could impact the service. It is anchored to an anycast address/Anycast-SID and hence the flow automatically reroutes on any ECMP-aware shortest-path to any other router part of the anycast set.

The two following sub-sections illustrate to traffic-engineering use-cases leveraging Anycast-SID.

4.1.1. Disjointness in dual-plane networks

Many networks are built according to the dual-plane design:
Each access region k is connected to the core by two C routers (C(1,k) and C(2,k)).

C(1,k) is part of plane 1 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) has a link to C(1, j)} iff {C(2,k) 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 (Cik-Cjk where i<>j) are set such that the route to an edge destination from a given plane stays within the plane unless the plane is partitioned.

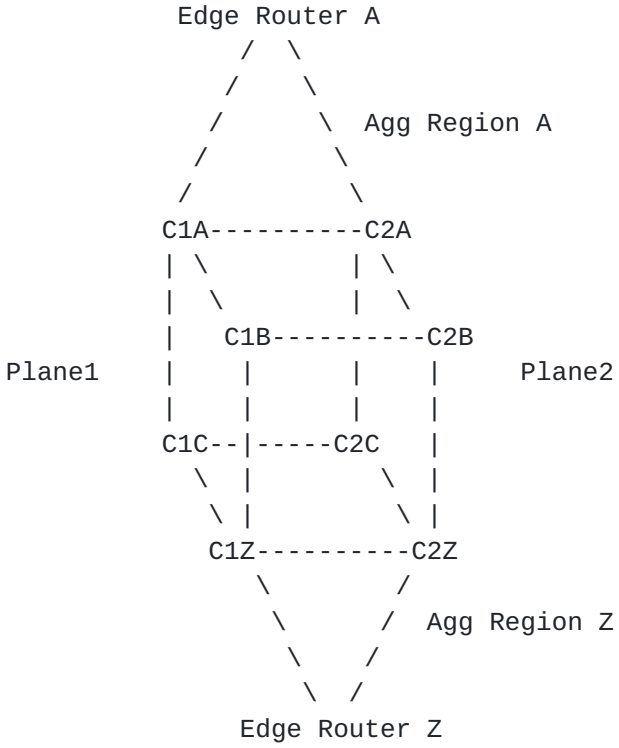


Figure 3: Dual-Plane Network and Disjointness

In the above network diagram, let us that the operator configures:

The four routers (C1A, C1B, C1C, C1Z) with an anycast loopback address 192.0.2.1/32 and an Anycast-SID 101.

The four routers (C2A, C2B, C2C, C2Z) with an anycast loopback address 192.0.2.2/32 and an Anycast-SID 102.

Edge router Z with Node-SID 109.

A can then use the three following segment lists to control its Z-destined traffic:

{109}: the traffic is load-balanced across any ECMP path through the network.

{101, 109}: the traffic is load-balanced across any ECMP path within the Plane1 of the network.

{102, 109}: the traffic is load-balanced across any ECMP path within the Plane2 of the network.

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

For example, let us assume a bank or a government customer has requested that the two flows F1 and F2 injected at A and destined to Z should be transported across disjoint paths. The operator could classify F1 (F2) at A and impose and SR header with the second (third) segment list. Focusing on F1 for the sake of illustration, A would route the packets based on the active segment, Anycast-SID 101, which steers the traffic along the ECMP-aware shortest-path to the closest router part of the Anycast-SID 101, C1A is this example. Once the packets have reached C1A, the second segment becomes active, Node-SID 109, which steers the traffic on the ECMP-aware shortest-path to Z. C1A load-balances the traffic between C1B-C1Z and C1C-C1Z and then C1Z forwards to Z.

This SR use-case has the following benefits:

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

Only two additional node segments (one Anycast-SID per plane).

ECMP-awareness.

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

4.1.2. CoS-based Traffic Engineering

Frequently, different classes of service need different path characteristics.

In the example below, a single-area international network with presence in four different regions of the world has lots of cheap network capacity from Region4 to Region1 via Region2 and some scarce expensive capacity via Region3.

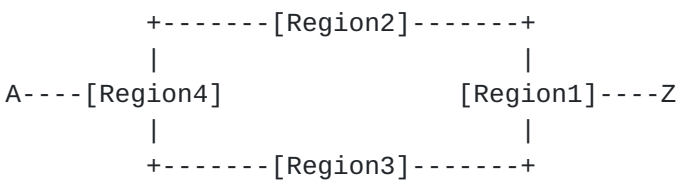


Figure 4: International Topology Example

In such case, the IGP metrics would be tuned to have a shortest-path from A to Z via Region2.

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

Let us illustrate how this can be solved with Segment Routing.

The operator would configure:
- All the core routers in Region3 with an anycast loopback 192.0.2.3/32 to which Anycast-SID 333 is attached.
- A loopback 192.0.2.9/32 on Z and would attach Node-SID 109 to it.
- The IGP metrics such that the shortest-path from Region4 to Region1 is via Region2, from Region4 to Region3 is directly to Region3, the shortest-path from Region3 to Region1 is not back via Region4 and Region2 but straight to Region1.

With this in mind, the operator would instruct A to apply the following policy for its Z-destined traffic:
- Voice traffic: impose segment-list {333, 109}
  - Anycast-SID 333 steers the Voice traffic along the ECMP-aware shortest-path to the closest core router in Region3, then Node-SID 109 steers the Voice traffic along the ECMP-aware shortest-path to Z. Hence the Voice traffic reaches Z from A via the low-latency path through Region3.

- Any other traffic: impose segment-list {109}: Node-SID 109 steers the Voice traffic along the ECMP-aware shortest-path to Z. Hence the bulk traffic reaches Z from A via the cheapest path for the operator.

This SR use-case has the following benefits:

Zero per-service state and signaling at midpoint and tailend nodes.

One additional anycast segment per region.

ECMP-awareness.

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

4.2. Distributed CSPF-based Traffic Engineering

In this section, we illustrate how a head-end router can map the result of its distributed CSPF computation into an SR segment list.

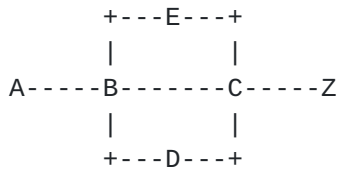


Figure 5: SRLG-based CSPF

Let us assume that in the above network diagram:

The operator configures a policy on A such that its Z-destined traffic must avoid SRLG1.

The operator configures SRLG1 on the link BC (or is learned dynamically from the IP/Optical interaction with the DWDM network).

The SRLG's are flooded in the link-state IGP.

The operator respectively configures the Node-SIDs 101, 102, 103, 104, 105 and 109 at nodes A, B, C, D, E and Z.

In that context, A can apply the following CSPF behavior:

- It prunes all the links affected by the SRLG1, computes an SPF on the remaining topology and picks one of the SPF paths.
  - In our example, A finds two possible paths ABECZ and ABDCZ and let's assume it takes the ABDCZ path.

- It translates the path as a list of segments
  - In our example, ABDCZ can be expressed as {104, 109}: a shortest path to node D, followed by a shortest-path to node Z.

- It monitors the status of the LSDB and upon any change impacting the policy, it either recomputes a path meeting the policy or update its translation as a list of segments.
  - For example, upon the loss of the link DC, the shortest-path to Z from D (Node-SID 109) goes via the undesired link BC. After a transient time immediately following such failure, the node A would figure out that the chosen path is no longer valid and instead select ABECZ which is translated as {103, 109}.

- This behavior is a local matter at node A and hence the details are outside the scope of this document.

The same use-case can be derived from any other C-SPF objective or constraint (TE affinity, TE latency, SRLG, etc.) as defined in [RFC5305] and [I-D.previdi-isis-te-metric-extensions]. Note that the bandwidth case is specific and hence is treated in Section 5.

4.3. Egress Peering Traffic Engineering

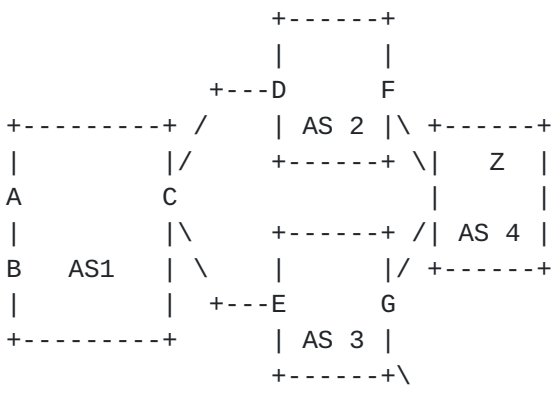


Figure 6: Egress peering traffic engineering

Let us assume that:

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

C sets next-hop-self before propagating the paths within AS1.

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

C only installs the path via AS2 in its RIB.

In that context, the operator of AS1 cannot apply the following traffic-engineering policy:

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.

This traffic-engineering policy can be supported thanks to the following SR configuration.

The operator configures:

C with a loopback 192.0.2.1/32 and attach the Node-SID 101 to it.

C to bind an external adjacency segment ([draft-filsfils-rtgwg-segment-routing-00]) to each of its peering interface.

For the sake of this illustration, let us assume that the external adjacency segments bound by C for its peering interfaces to (D, AS2) and (E, AS3) are respectively 9001 and 9002.

These external adjacencies (and their attached segments) are flooded within the IGP domain of AS1 [RFC5316].

As a result, the following information is available within AS1:
ISIS Link State Database:

- Node-SID 101 is attached to IP address 192.0.2.1/32 advertised by C.
- C is connected to a peer D with external adjacency segment 9001.
- C is connected to a peer E with external adjacency segment 9002.

BGP Database:

- Z is reachable via 192.0.2.1 with AS Path {AS2, AS4}.
- Z is reachable via 192.0.2.1 with AS Path {AS3, AS4}.

The operator of AS1 can thus meet its traffic-engineering objective by enforcing the following policies:

A should apply the segment list {101, 9001} to 60% of the Z-destined traffic and the segment list {101, 9002} to the rest.

B should apply the segment list {101, 9001} to 80% of the Z-destined traffic and the segment list {101, 9002} to the rest.

Node segment 101 steers the traffic to C.

External adjacency segment 9001 forces the traffic from C to (D, AS2), without any IP lookup at C.

External adjacency segment 9002 forces the traffic from C to (E, AS3), without any IP lookup at C.

A and B can also use the described segments to assess the liveness of the remote peering links, see OAM section.

4.4. Deterministic non-ECMP Path

The previous sections have illustrated the ability to steer traffic along ECMP-aware shortest-paths. SR is also able to express deterministic non-ECMP path: i.e. as a list of adjacency segments. We illustrate such an use-case in this section.

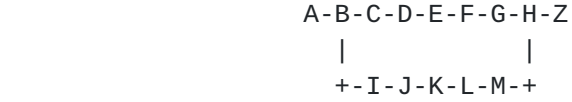


Figure 7: Non-ECMP deterministic path

In the above figure, it is assumed all nodes are SR capable and only the following SIDs are advertised:

- A advertises Adj-SID 9001 for its adjacency to B
- B advertises Adj-SID 9002 for its adjacency to C
- C advertises Adj-SID 9003 for its adjacency to D
- D advertises Adj-SID 9004 for its adjacency to E
- E advertises Adj-SID 9001 for its adjacency to F
- F advertises Adj-SID 9002 for its adjacency to G
- G advertises Adj-SID 9003 for its adjacency to H
- H advertises Adj-SID 9004 for its adjacency to Z
- E advertises Node-SID 101
- Z advertises Node-SID 109

The operator can steer the traffic from A to Z via a specific non-ECMP path ABCDEFGHZ by imposing the segment list {9001, 9002, 9003, 9004, 9001, 9002, 9003, 9004}.

The following sub-sections illustrate how the segment list can be compressed.

4.4.1. Node Segment

Clearly the same exact path can be expressed with a two-entry segment list {101, 109}.

This example illustrates that a Node Segment can also be used to express deterministic non-ECMP path.

4.4.2. Forwarding Adjacency

The operator can configure Node B to create a forwarding-adjacency to node H along an explicit path BCDEFGH. The following behaviors can then be automated by B:

B attaches an Adj-SID (e.g. 9007) to that forwarding adjacency together with an ERO sub-sub-TLV which describes the explicit path BCDEFGH.

B installs in its Segment Routing Database the following entry:

Active segment: 9007.

Operation: NEXT and PUSH {9002, 9003, 9004, 9001, 9002, 9003}

As a result, the operator can configure node A with the following compressed segment list {9001, 9007, 9004}.

4.5. Load-balancing among non-parallel links

A given node may assign the same Adj-SID to multiple of its adjacencies, even if these ones lead to different neighbors. This may be useful to support traffic engineering policies.

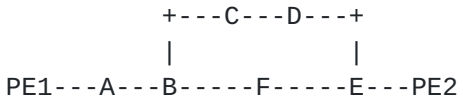


Figure 8: Adj-SID For Multiple (non-parallel) Adjacencies

In the above example, let us assume that the operator:

Requires PE1 to load-balance its PE2-destined traffic between the ABCDE and ABFE paths.

Configures B with Node-SID 102 and E with Node-SID 202.

Configures B to advertise an individual Adj-SID per adjacency (e.g. 9001 for BC and 9002 for BF) and, in addition, an Adj-SID for the adjacency set (BC, BF) (e.g. 9003).

With this context in mind, the operator achieves its objective by configuring the following traffic-engineering policy at PE1 for the PE2-destined traffic: {102, 9001, 202}:

Node-SID 102 steers the traffic to B.

Adj-SID 9003 load-balances the traffic to C or F.

From either C or F, Node-SID 202 steers the traffic to PE2.

In conclusion, the traffic is load-balanced between the ABCDE and ABFE paths, as desired.

5. 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.

The first sub-section details the capacity planning process and the role of ECMP load-balancing. We highlight the relevance of SR in that context.

The next two sub-sections document two use-cases of SR-based traffic engineering with bandwidth admission control.

The second sub-section documents a concrete SR applicability involving centralized-based admission control. This is often referred to as the "SDN/SR use-case".

The third sub-section introduces a future research topic involving the notion of residual bandwidth introduced in [I-D.atlas-mpls-te-express-path].

5.1. 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.

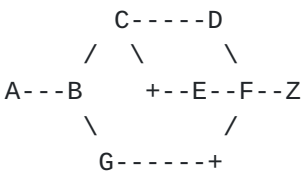


Figure 9: Capacity Planning an ECMP-based demand

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.

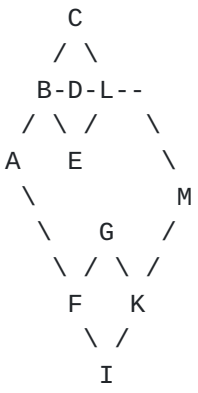


Figure 10: ECMP Topology Example

Segment Routing offers a simple support for such ECMP-based shortest-

path placement: a node segment. A single node segment enumerates all the ECMP paths along the shortest-path.

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.

Segment Routing offers a simple support for explicit path policy. Let us provide two examples based on Figure 10.

First example: let us assume that the process has selected the flow AM for traffic-engineering away from its ECMP-enabled shortest path and flow AM must avoid consuming resources on the LM and the FG links.

The solution is straightforward: A sends its M-destined traffic towards the nhop F with a two-label stack where the top label is the adjacent segment FI and the next label is the node segment to M. Alternatively, a three-label stack with adjacency segments FI, IK and KM could have been used.

Second example: let us assume that AM is still the selected flow but the constraint is relaxed to only avoid using resources from the LM link.

The solution is straightforward: A sends its M-destined traffic towards the nhop F with a one-label stack where the label is the node segment to M. Note that while the AM flow has been traffic-engineered away from its natural shortest-path (ECMP across three paths), the traffic-engineered path is still ECMP-aware and leverages two of the three initial paths. This is accomplished with a single-label stack and without the enumeration of one tunnel per path.

Under the light of these examples, Segment Routing offers an interesting solution for Capacity Planning because:

One node segment represents the set of ECMP-aware shortest paths.

Adjacency segments allow to express any explicit path.

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

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

The explicitly-engineered traffic (which is a minority) still benefits from the ECMP-awareness of the node segments within their segment list.

Only the head-end of a traffic-engineering policy maintains state. The midpoints and tail-ends do not maintain any state.

5.2. SDN /SR use-case

The heart of the application of SR to the SDN use-case lies in the SDN controller, also called Stateful PCE ([I-D.ietf-pce-stateful-pce]).

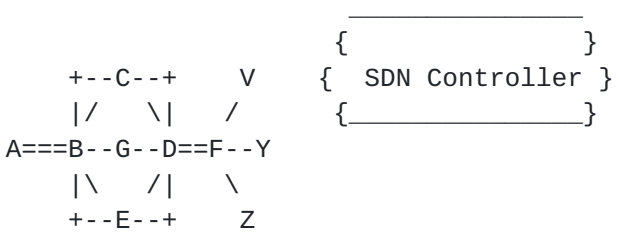
The SDN controller is responsible to control the evolution of the traffic matrix and topology. It accepts or denies the addition of new traffic into the network. It decides how to route the accepted traffic. It monitors the topology and upon failure, determines the minimum traffic that should be rerouted 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. [RFC5101] 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.ward-i2rs-framework], [I-D.crabbe-pce-pce-initiated-lsp] and [draft-msiva-pce-pcep-segment-routing-extensions-00]).

5.2.1. Illustration



SDN/SR use-case

Let us assume that in the above network diagram:

An SDN Controller (SC) is connected to the network and is able to retrieve the topology and traffic information, as well as set traffic-engineering policies on the network nodes.

The operator (likely via the SDN Controller) as provisioned the Node-SIDs 101, 102, 103, 104, 105, 106, 107, 201, 202 and 203 respectively at nodes A, B, C, D, E, F, G, V, Y and Z.

All the links have the same BW (e.g. 10G) and IGP cost (e.g. 10) except the links BG and GD which have IGP cost 50.

Each described node connectivity is formed as a bundle of two links, except (B, G) and (G, D) which are formed by a single link each.

Flow FV is traveling from A to destinations behind V.

Flow FY is traveling from A to destinations behind Y.

Flow FZ is traveling from A to destinations behind Z.

The SDN Controller has admitted all these flows and has let A apply the default SR policy: "map a flow onto its ECMP-aware shortest-path".

In this example, this means that A respectively maps the flows FV onto segment list {201}, FY onto segment list {202} and FZ onto segment list {203}.

In this example, the reader should note that the SDN Controller knows what A would do and hence knows and controls that none of these flows are mapped through G.

Let us describe what happens upon the failure of one of the two links E-D.

The SDN Controller monitors the link-state database and detects a congestion risk due to the reduced capacity between E and D. Specifically, SC updates its simulation of the traffic according to the policies he instructed the network to use and discovers that too much traffic is mapped on the remaining link E-D.

The SDN Controller then computes the minimum number of flows that should be deviated from their existing path. For example, let us assume that the flow FZ is selected.

The SDN controller then computes an explicit path for this flow. For example, let us assume that the chosen path is ABGDFZ.

The SDN controller then maps the chosen path into an SR-based policy. In our example, the path ABGDFZ is translated into a segment list {107, 203}. Node-SID steers the traffic along ABG and then Node-SID 203 steers the traffic along GDFZ.

The SDN controller then applies the following traffic-engineering policy at A: "map any packet of the classified flow FZ onto segment-list {107, 203}". The SDN Controller uses PCEP extensions to instantiate that policy at A ([draft-msiva-pce-pcep-segment-routing-extensions-00]).

As soon as A receives the PCEP message, it enforces the policy and the traffic classified as FZ is immediately mapped onto segment list {107, 203}.

This immediately eliminate the congestion risk. Flows FV and FY were untouched and keep using the ECMP-aware shortest-path. The minimum amount of traffic was rerouted (FZ). No signaling hop-by-hop through the network from A to Z is required. No admission control hop-by-hop is required. No state needs to be maintained by B, G, D, F or Z. The only maintained state is within the SDN controller and the head-end node (A).

5.2.2. Benefits

In the context of Centralized-Based Optimization and the SDN use-case, here are the benefits provided by the SR architecture:

- 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 (Section 3.

Optimum virtualization: the policy state is in the packet header and not in the intermediate node along the policy. 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 lost programming the midpoints and tail-end along the policy.

5.2.3. Dataset analysis

A future version of this document will report some analysis of the application of the SDN/SR use-case to real operator data sets.

A first, incomplete, report is available here below.

5.2.3.1. Example 1

The first 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 riding on 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 riding off its shortest path.

We applied the SDN/SR use-case to this dataset. 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 SR segment-lists.

As a result, we measured that 98% of the traffic would be kept on its normal policy (ride 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 SR-capable ingress edge node maps the traffic on a single Node-ID 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.

5.3. Residual Bandwidth

The notion of Residual Bandwidth (RBW) is introduced by [I-D.atlas-mpls-te-express-path].

A future version of this document will describe the SR/RBW research opportunity.

6. SR co-existence and interworking with other MPLS Control Plane

The first section describes the co-existence of SR with other MPLS Control Plane. The second section documents a method to migrate from LDP to SR-based MPLS tunneling. The third section documents the interworking of LDP and SR in the case of non-homogenous deployment. The fourth use-case describes how a partial SR deployment can be used to provide SR benefits to LDP-based traffic. The fifth section describes a possible application of SR in the context of inter-domain MPLS use-cases.

6.1. Ship-in-the-night coexistence

We call "MPLS Control Plane Client (MCC)" any control-plane protocol installing forwarding entries in the MPLS dataplane. SR, LDP, RSVP-TE, BGP 3107, VPNv4, etc. are examples of MCCs.

An MCC, operating at node N, must ensure that the incoming label it installs in the MPLS dataplane of Node N has been uniquely allocated to himself.

Thanks to the defined segment allocation rule and specifically the notion of the SRGB, SR can co-exist with any other MCC.

This is clearly the case for the adjacency segment: it is a local label allocated by the label manager, as for any MCC.

This is clearly the case for the prefix segment: the label manager allocates the SRGB set of labels to the SR MCC client and the operator ensures the unique allocation of each global prefix segment/ label within the allocated SRGB set.

Note that this static label allocation capability of the label manager has been existing for many years across several vendors and hence is not new. Furthermore, note that the label-manager ability to statically allocate a range of labels to a specific application is not new either. This is required for MPLS-TP operation. In this case, the range is reserved by the label manager and it is the MPLS-TP NMS (acting as an MCC) that ensures the unique allocation of any label within the allocated range and the creation of the related MPLS forwarding entry.

Let us illustrate an example of ship-in-the-night (SIN) coexistence.

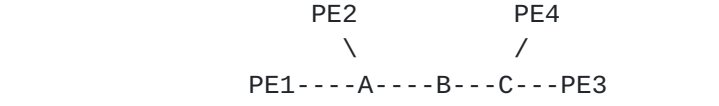


Figure 11: SIN coexistence

The EVEN VPN service is supported by PE2 and PE4 while the ODD VPN service is supported by PE1 and PE3. The operator wants to tunnel the ODD service via LDP and the EVEN service via SR.

This can be achieved in the following manner:

The operator configures PE1, PE2, PE3, PE4 with respective loopbacks 192.0.2.201/32, 192.0.2.202/32, 192.0.2.203/32, 192.0.2.204/32. These PE's advertised their VPN routes with next-hop set on their respective loopback address.

The operator configures A, B, C with respective loopbacks 192.0.2.1/32, 192.0.2.2/32, 192.0.2.3/32.

The operator configures PE2, A, B, C and PE4 with SRGB {100, 300}.

The operator attaches the respective Node-SIDs 202, 101, 102, 103 and 204 to the loopbacks of nodes PE2, A, B, C and PE4. The Node-SID's are configured to request penultimate-hop-popping.

PE1, A, B, C and PE3 are LDP capable.

PE1 and PE3 are not SR capable.

PE3 sends an ODD VPN route to PE1 with next-hop 192.0.2.203 and VPN label 10001.

From an LDP viewpoint: PE1 received an LDP label binding (1037) for FEC 192.0.2.203/32 from its nhop A. A received an LDP label binding (2048) for that FEC from its nhop B. B received an LDP label binding (3059) for that FEC from its nhop C. C received implicit-null LDP binding from its next-hop PE3.

As a result, PE1 sends its traffic to the ODD service route advertised by PE3 to next-hop A with two labels: the top label is 1037 and the bottom label is 10001. A swaps 1037 with 2048 and forwards to B. B swaps 2048 with 3059 and forwards to C. C pops 3059 and forwards to PE3.

PE4 sends an EVEN VPN route to PE2 with next-hop 192.0.2.204 and VPN label 10002.

From an SR viewpoint: PE1 maps the IGP route 192.0.2.204/32 onto Node-SID 204; A swaps 204 with 204 and forwards to B; B swaps 204 with 204 and forwards to C; C pops 204 and forwards to PE4.

As a result, PE2 sends its traffic to the VPN service route advertised by PE4 to next-hop A with two labels: the top label is 204 and the bottom label is 10002. A swaps 204 with 204 and forwards to B. B swaps 204 with 204 and forwards to C. C pops 204 and forwards to PE4.

The two modes of MPLS tunneling co-exist.

The ODD service is tunneled from PE1 to PE3 through a continuous LDP LSP traversing A, B and C.

The EVEN service is tunneled from PE2 to PE4 through a continuous SR node segment traversing A, B and C.

6.1.1. MPLS2MPLS co-existence

We want to highlight that several MPLS2MPLS entries can be installed in the dataplane for the same prefix.

Let us examine A's MPLS forwarding table as an example:

Incoming label: 1037
- outgoing label: 2048
- outgoing nhop: B
- Note: this entry is programmed by LDP for 192.0.2.203/32

Incoming label: 203
- outgoing label: 203
- outgoing nhop: B
- Note: this entry is programmed by SR for 192.0.2.203/32

These two entries can co-exist because their incoming label is unique. The uniqueness is guaranteed by the label manager allocation rules.

The same applies for the MPLS2IP forwarding entries.

6.1.2. IP2MPLS co-existence

By default, we propose that if both LDP and SR propose an IP2MPLS entry for the same IP prefix, then the LDP route is selected.

A local policy on a router MUST allow to prefer the SR-provided IP2MPLS entry.

6.2. Migration from LDP to SR

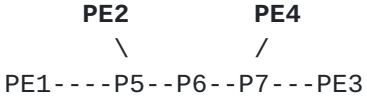


Figure 12: Migration

Several migration techniques are possible. We describe one technique inspired by the commonly used method to migrate from one IGP to another.

T0: all the routers run LDP. Any service is tunneled from an ingress PE to an egress PE over a continuous LDP LSP.

T1: all the routers are upgraded to SR. They are configured with the SRGB range (100, 200). PE1, PE2, PE3, PE4, P5, P6 and P7 are respectively configured with the node segments 101, 102, 103, 104,

105, 106 and 107 (attached to their service-recursing loopback).

At this time, the service traffic is still tunneled over LDP LSP. For example, PE1 has an SR node segment to PE3 and an LDP LSP to PE3 but by default, as seen earlier, the LDP IP2MPLS encapsulation is preferred.

T2: the operator enables the local policy at PE1 to prefer SR IP2MPLS encapsulation over LDP IP2MPLS.

The service from PE1 to any other PE is now riding over SR. All other service traffic is still transported over LDP LSP.

T3: gradually, the operator enables the preference for SR IP2MPLS encapsulation across all the edge routers.

All the service traffic is now transported over SR. LDP is still operational and services could be reverted to LDP.

T4: LDP is unconfigured from all routers.

6.3. SR and LDP Interworking

In this section, we analyze a use-case where SR is available in one part of the network and LDP is available in another part. We describe how a continuous MPLS tunnel can be built throughout the network.

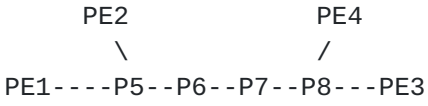


Figure 13: SR and LDP Interworking

Let us analyze the following example:

P6, P7, P8, PE4 and PE3 are LDP capable.

PE1, PE2, P5 and P6 are SR capable. PE1, PE2, P5 and P6 are configured with SRGB (100, 200) and respectively with node segments 101, 102, 105 and 106.

A service flow must be tunneled from PE1 to PE3 over a continuous MPLS tunnel encapsulation. We need SR and LDP to interwork.

6.3.1. LDP to SR

In this section, we analyze a right-to-left traffic flow.

PE3 has learned a service route whose nhop is PE1. PE3 has an LDP label binding from the nhop P8 for the FEC "PE1". Hence PE3 sends its service packet to P8 as per classic LDP behavior.

P8 has an LDP label binding from its nhop P7 for the FEC "PE1" and hence P8 forwards to P7 as per classic LDP behavior.

P7 has an LDP label binding from its nhop P6 for the FEC "PE1" and hence P7 forwards to P6 as per classic LDP behavior.

P6 does not have an LDP binding from its nhop P5 for the FEC "PE1". However P6 has an SR node segment to the IGP route "PE1". Hence, P6 forwards the packet to P5 and swaps its local LDP-label for FEC "PE1" by the equivalent node segment (i.e. 101).

P5 pops 101 (assuming PE1 advertised its node segment 101 with the penultimate-pop flag set) and forwards to PE1.

PE1 receives the tunneled packet and processes the service label.

The end-to-end MPLS tunnel is built from an LDP LSP from PE3 to P6 and the related node segment from P6 to PE1.

6.3.2. SR to LDP

In this section, we analyze the left-to-right traffic flow.

We assume that the operator configures P5 to act as a Segment Routing Mapping Server (SRMS) and advertise the following mappings: (P7, 107), (P8, 108), (PE3, 103) and (PE4, 104).

The mappings advertised by an SR mapping server result from local policy information configured by the operator. IF PE3 had been SR capable, the operator would have configured PE3 with node segment 103. Instead, as PE3 is not SR capable, the operator configures that policy at the SRMS and it is the latter which advertises the mapping. Multiple SRMS servers can be provisioned in a network for redundancy.

The mapping server advertisements are only understood by the SR capable routers. The SR capable routers install the related node segments in the MPLS dataplane exactly like if the node segments had been advertised by the nodes themselves.

For example, PE1 installs the node segment 103 with nhop P5 exactly

as if PE3 had advertised node segment 103.

PE1 has a service route whose nhop is PE3. PE1 has a node segment for that IGP route: 103 with nhop P5. Hence PE1 sends its service packet to P5 with two labels: the bottom label is the service label and the top label is 103.

P5 swaps 103 for 103 and forwards to P6.

P6's next-hop for the IGP route "PE3" is not SR capable (P7 does not advertise the SR capability). However, P6 has an LDP label binding from that next-hop for the same FEC (e.g. LDP label 1037). Hence, P6 swaps 103 for 1037 and forwards to P7.

P7 swaps this label with the LDP-label received from P8 and forwards to P8.

P8 pops the LDP label and forwards to PE3.

PE3 receives the tunneled packet and processes the service label.

The end-to-end MPLS tunnel is built from an SR node segment from PE1 to P6 and an LDP LSP from P6 to PE3.

6.4. Leveraging SR benefits for LDP-based traffic

SR can be deployed such as to enhance LDP transport. The SR deployment can be limited to the network region where the SR benefits are most desired.

In Figure 14, let us assume:

All link costs are 10 except FG which is 30.

All routers are LDP capable.

X, Y and Z are PE's participating to an important service S.

The operator requires 50msec link-based FRR for service S.

A, B, C, D, E, F and G are SR capable.

X, Y, Z are not SR capable, e.g. as part of a staged migration from LDP to SR, the operator deploys SR first in a sub-part of the network and then everywhere.

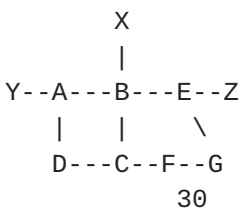


Figure 14: Leveraging SR benefits for LDP-based-traffic

The operator would like to resolve the following issues:

To protect the link BA along the shortest-path of the important flow XY, B requires an RLFA repair tunnel to D and hence a directed LDP session from B to D. The operator does not like these dynamically established multi-hop LDP sessions and would seek to eliminate them.

There is no LFA/RLFA solution to protect the link BE along the shortest path of the important flow XZ. The operator wants a guaranteed link-based FRR solution.

The operator can meet these objectives by deploying SR only on A, B, C, D, E and F:

The operator configures A, B, C, D, E, F and G with SRGB (100, 200) and respective node segments 101, 102, 103, 104, 105, 106 and 107.

The operator configures D as an SR Mapping Server with the following policy mapping: (X, 201), (Y, 202), (Z, 203}.

Each SR node automatically advertises local adjacency segment for its IGP adjacencies. Specifically, F advertises adjacency segment 9001 for its adjacency FG.

A, B, C, D, E, F and G keep their LDP capability and hence the flows XY and XZ are transported over end-to-end LDP LSP's.

For example, LDP at B installs the following MPLS dataplane entries:
Incoming label: local LDB label bound by B for FEC Y
Outgoing label: LDP label bound by A for FEC Y
Outgoing nhop: A
Incoming label: local LDB label bound by B for FEC Z
Outgoing label: LDP label bound by E for FEC Z
Outgoing nhop: E

The novelty comes from how the backup chains are computed for these LDP-based entries. While LDP labels are used for the primary nhop

and outgoing labels, SR information is used for the FRR construction. In steady state, the traffic is transported over LDP LSP. In transient FRR state, the traffic is backup thanks to the SR enhanced capabilities.

This helps meet the requirements of the operator:

Eliminate directed LDP session.

Guaranteed FRR coverage.

Keep the traffic over LDP LSP in steady state.

Partial SR deployment only where needed.

6.4.1. Eliminating Directed LDP Session

B's MPLS entry to Y becomes:
- Incoming label: local LDB label bound by B for FEC Y
  Outgoing label: LDP label bound by A for FEC Y
  Backup outgoing label: SR node segment for Y {202}
  Outgoing nhop: A
  Backup nhop: repair tunnel: node segment to D {104}
  with outgoing nhop: C

In steady-state, X sends its Y-destined traffic to B with a top label which is the LDP label bound by B for FEC Y. B swaps that top label for the LDP label bound by A for FEC Y and forwards to A. A pops the LDP label and forwards to Y.

Upon failure of the link BA, B swaps the incoming top-label with the node segment for Y (202) and sends the packet onto a repair tunnel to D (node segment 104). Thus, B sends the packet to C with the label stack {104, 202}. C pops the node segment 104 and forwards to D. D swaps 202 for 202 and forwards to A. A's nhop to Y is not SR capable and hence A swaps the incoming node segment 202 to the LDP label announced by its next-hop (in this case, implicit null).

After IGP convergence, B's MPLS entry to Y will become:
- Incoming label: local LDB label bound by B for FEC Y
  Outgoing label: LDP label bound by C for FEC Y
  Outgoing nhop: C

And the traffic XY travels again over the LDP LSP.

Conclusion: the operator has eliminated its first problem: directed LDP sessions are no longer required and the steady-state traffic is still transported over LDP. The SR deployment is confined to the

area where these benefits are required.

6.4.2. Guaranteed FRR coverage

B's MPLS entry to Z becomes:
- Incoming label: local LDB label bound by B for FEC Z
  Outgoing label: LDP label bound by E for FEC Z
  Backup outgoing label: SR node segment for Z {203}
  Outgoing nhop: E
  Backup nhop: repair tunnel to G: {106, 9001}

    G is reachable from B via the combination of a node segment to F {106} and an adjacency segment FG {9001}

    Note that {106, 107} would have equally work. Indeed, in many case, P's shortest path to Q is over the link PQ. The adjacency segment from P to Q is required only in very rare topologies where the shortest-path from P to Q is not via the link PQ.

In steady-state, X sends its Z-destined traffic to B with a top label which is the LDP label bound by B for FEC Z. B swaps that top label for the LDP label bound by E for FEC Z and forwards to E. E pops the LDP label and forwards to Z.

Upon failure of the link BE, B swaps the incoming top-label with the node segment for Z (203) and sends the packet onto a repair tunnel to G (node segment 106 followed by adjacency segment 9001). Thus, B sends the packet to C with the label stack {106, 9001, 203}. C pops the node segment 106 and forwards to F. F pops the adjacency segment 9001 and forwards to G. G swaps 203 for 203 and forwards to E. E's nhop to Z is not SR capable and hence E swaps the incoming node segment 203 for the LDP label announced by its next-hop (in this case, implicit null).

After IGP convergence, B's MPLS entry to Z will become:
- Incoming label: local LDB label bound by B for FEC Z
  Outgoing label: LDP label bound by C for FEC Z
  Outgoing nhop: C

And the traffic XZ travels again over the LDP LSP.

Conclusion: the operator has eliminated its second problem: guaranteed FRR coverage is provided. The steady-state traffic is still transported over LDP. The SR deployment is confined to the area where these benefits are required.

6.5. Inter-AS Option C, Carrier's Carrier and Seamless MPLS

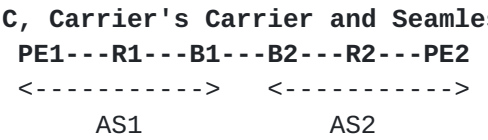


Figure 15: Inter-AS Option C

In Inter-AS Option C [RFC4364], B2 advertises to B1 a BGP3107 route for PE2 and B1 reflects it to its internal peers, such as PE1. PE1 learns from a service route reflector a service route whose nhop is PE2. PE1 resolves that service route on the BGP3107 route to PE2. That BGP3107 route to PE2 is itself resolved on the AS1 IGP route to B1.

If AS1 operates SR, then the tunnel from PE1 to B1 is provided by the node segment from PE1 to B1.

PE1 sends a service packet with three labels: the top one is the node segment to B1, the next-one is the BGP3107 label provided by B1 for the route "PE2" and the bottom one is the service label allocated by PE2.

The same straightforward SR applicability is derived for CsC and Seamless MPLS ([I-D.ietf-mpls-seamless-mpls]).

7. OAM

7.1. Monitoring a remote bundle

This section documents a few representative SR/OAM use-cases.

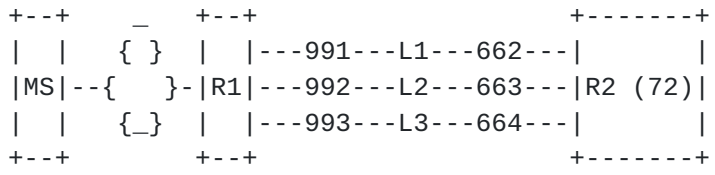


Figure 16: Probing all the links of a remote bundle

In the above figure, a monitoring system (MS) 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 segment information from the IGP LSDB and appends the following segment list: {72, 662, 992, 664} on its IP probe (whose source and destination addresses are the address of AA).

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 SR-based probe to the first SR router. The SR domain forwards the probe to R2 (72 is the node segment of R2). R2 forwards the probe to R1 over link L1 (adjacency segment 662). R1 forwards the probe to R2 over link L2 (adjacency segment 992). R2 forwards the probe to R1 over link L3 (adjacency segment 664). R1 then forwards the IP probe to AA as per classic IP forwarding.

7.2. Monitoring a remote peering link

In Figure 6, node A can monitor the dataplane liveness of the unidirectional peering link from C to D of AS2 by sending an IP probe with destination address A and segment list {101, 9001}. Node-SID 101 steers the probe to C and External Adj-SID 9001 steers the probe from C over the desired peering link to D of AS2. The SR header is removed by C and D receives a plain IP packet with destination address A. D returns the probe to A through classic IP forwarding. BFD Echo mode ([RFC5880]) would support such liveliness unidirectional link probing application.

8. IANA Considerations

TBD

9. Manageability Considerations

TBD

10. Security Considerations

TBD

11. Acknowledgements

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

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.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, December 2006.

[RFC5101] Claise, B., "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information", RFC 5101, January 2008.

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

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

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

12.2. Informative References

[I-D.atlas-mpls-te-express-path]
Atlas, A., Drake, J., Giacalone, S., Ward, D., Previdi, S., and C. Filsfils, "Performance-based Path Selection for Explicitly Routed LSPs", draft-atlas-mpls-te-express-path-02 (work in progress), February 2013.

[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-01 (work in progress), April 2013.

[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-03 (work in progress), May 2013.

[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-03 (work in progress), May 2013.

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-04 (work in progress), May 2013.

[I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-02 (work in progress), May 2013.

[I-D.previdi-isis-te-metric-extensions]
Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions-03 (work in progress), February 2013.

[I-D.ward-i2rs-framework]
Atlas, A., Nadeau, T., and D. Ward, "Interface to the Routing System Framework", draft-ward-i2rs-framework-00 (work in progress), February 2013.

[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.

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

[draft-filsfils-rtgwg-segment-routing-00]
Filsfils, C. and S. Previdi, "Segment Routing Architecture", June 2013.

[draft-msiva-pce-pcep-segment-routing-extensions-00]
Filsfils, C. and S. Sivabalan, "PCEP Extensions for Segment Routing", May 2013.

[draft-previdi-isis-segment-routing-extensions-00]
Previdi, S., Filsfils, C., and A. Bashandy, "IS-IS Segment

Routing Extensions", May 2013.

[draft-psenak-ospf-segment-routing-extensions-00]
Psenak, P. and S. Previdi, "OSPF Segment Routing
Extensions", May 2013.

[draft-rtgwg-bgp-pic-01.txt]
Filsfils, C., Bashandy, A., and P. Mohapatra, "BGP Prefix
Independent Convergence", March 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