SPRING Working Group P. Sarkar, Ed.
Internet-Draft H. Gredler
Intended status: Standards Track Juniper Networks, Inc.
Expires: January 7, 2016 July 6, 2015

# Anycast Segments in MPLS based SPRING
# draft-psarkar-spring-mpls-anycast-segments-00

## Abstract

Instead of forwarding to a specific device or to all devices in a group, anycast addresses, let network devices forward a packet to (or steer it through) one or more topologically nearest devices in a specific group of network devices. [I-D.ietf-spring-segment-routing] extended the use of anycast addresses to a SPRING network, wherein a group of SPRING-capable devices can represent a anycast address, by having the same SRGB label block provisioned on all the devices and each one of them advertising the same anycast prefix segment (or Anycast SID).

This document describes a proposal for implementing anycast prefix segments in SPRING, without the need to have the same SRGB block (label ranges) provisioned across all the member devices in the group. Each node can be provisioned with a separate SRGB from the label range supported by the specfic hardware platform.

## 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 7, 2016.

Copyright Notice

Copyright (c) 2015 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

Anycast is a network addressing scheme and routing methodology in which packets from a single source device are forwarded to the topologically nearest node in a group of potential receiving devices, all identified by the same anycast address. There are various useful usecases of anycast addresses, and discussion of the same are outside the scope of this document.

[I-D.ietf-spring-segment-routing] extended the use of anycast addresses to SPRING networks. An operator may combine a group of SPRING-enabled nodes to form a anycast group, by picking a anycast

address and a segment identifier (hereon referred to as SID) to represent the group, and then provisioning all the nodes with the same address and SID. Once provisioned, each device in the group advertises the corresponding anycast address in it's IGP link-state advertisements along with the SID provisioned. Source devices on receiving such anycast prefix segment advertisements, finds out the topologically nearest device that originated the anycast segment and forwards packets destined to the same on the shortest-path to the nearest device.

[I-D.ietf-spring-segment-routing] also requires all devices in a given anycast group to implement the exact same SRGB block. While this requirement will always be met in SPRING network deployed over IPV6 forwarding plane [I-D.previdi-6man-segment-routing-header], the same may not be easily met in all SPRING deployments over MPLS dataplane [I-D.ietf-spring-segment-routing-mpls].

In MPLS-based SPRING deployments the segments on a given source router are actually mapped to a MPLS labels allocated from the local label pool carved out by the device for accomodating the SRGB block. In multi-vendor deployments with various types of devices deployed in the same network topology, such a anycast group may contain a good combination of devices from different vendors and have different internal hardware capabilities. In such environments it is not sufficient to assume that all the devices in a anycast group will be able to allocate exactly the same range of labels for implementing the SRGB. In reality, getting a common range of labels among all the various vendors is not feasible.

This documents provides mechanisms to implement a anycast segments with any kind of device in a multi-vendor netwrok deployment without requiring to provision the same exact range of labels for SRGB on all the devices.

## 2. Problem Statement

To better illustrate the problem let us consider an example topology using anycast segments as shown in Figure 1 below.

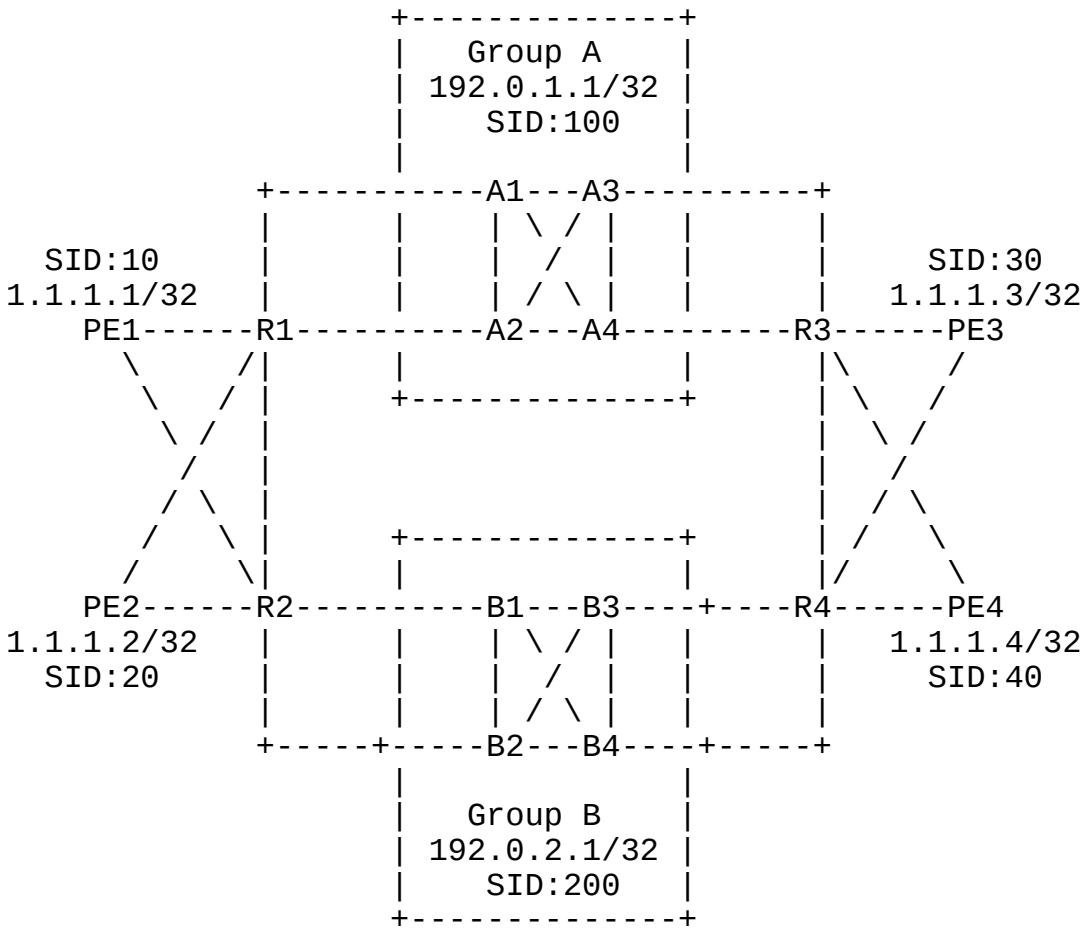


Figure 1: Topology 1

In Figure 1 above, there are two groups of transit devices. Group A consists of devices {A1, A2, A3 and A4}. They are all provisioned with the anycast address 192.0.1.1/32 and the anycast SID 100. Similarly, group B consists of devices {B1, B2, B3 and B4} and are all provisioned with the anycast address 192.0.1.2/32, anycast SID 200. In the above network topology, each PE device is connected to two routers in each of the groups A and B.

Following are all the possible ECMP paths between the various pairs of PE devices.

- P1: via {R1, A1, A3, R3}
- P2: via {R1, A1, A4, R3}
- P3: via {R1, A2, A3, R3}

- P4: via {R1, A2, A4, R3}
- P5: via {R2, B1, B3, R4}
- P6: via {R2, B1, B4, R4}
- P7: via {R2, B2, B3, R4}
- P8: via {R2, B2, B4, R4}

As seen above, there is always eight ECMP paths between each of pair of PE devices. The network operator may not wish to utilize all possible ECMP paths for all possible types of traffic flowing between a given pair of PE devices. It may be more useful for use paths P1, P2, P3 and P4 for certain types of traffic and use paths P5, P6, P7 and P8 for all other types of traffic between the same PE devices. If so desired, operators may use these anycast groups A and B and the corresponding anycast segment to impose a segment-list to forward the respective traffic flows over the desired specific paths as shown below. Figure 2 below depicts a expanded view of the paths via group A. The range labels allocated for SRGB on each of the devices in group A are also mentioned in this diagram.

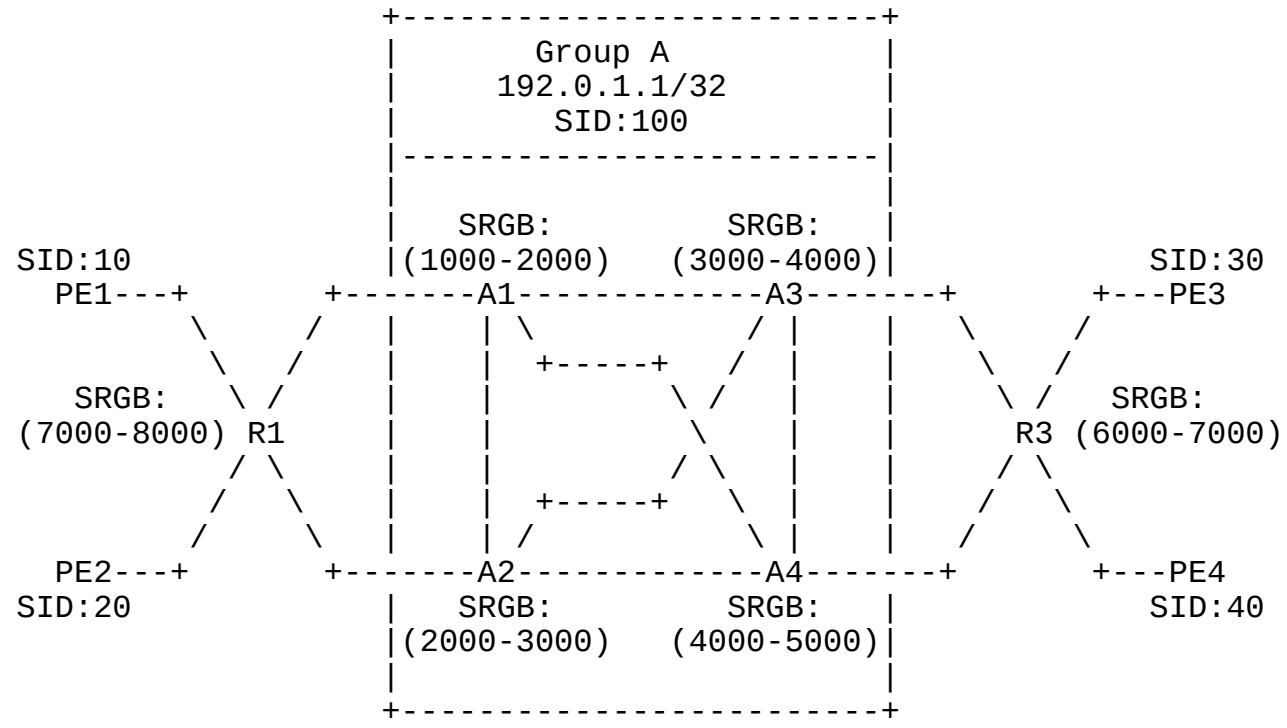


Figure 2: Transit paths via anycast group A

In the above topology, if device PE1 (or PE2) requires to send a packet to the device PE3 (or PE4) it needs to encapsulate the packet in a MPLS payload with the following stack of labels.

- Label allocated R1 for anycast SID 100 (outer label)
- Label allocated by the nearest router in group A for SID 30 (for destination PE3)

While the first label is easy to compute, in this case since there are more than one topologically nearest devices (A1 and A2), unless A1 and A2 implement same exact SRGB, determining the second label is impossible. In all likeness, devices A1 and A2 may be devices from different hardware vendors and it may not implement the same exact SRGB label ranges. In such cases, separate labels are allocated by A1 and A2 (1030 and 2030 respectively, in the above example). Hence, PE1 (or PE2) cannot compute an appropriate label stack to steer the packet exclusively through the group A devices. Same holds true for devices PE3 and PE4 when trying to send a packet to PE1 or PE2.

# 3. Solution

## 3.1. Anycast Segment Label

This document introduces the term 'Anycast Segment Label' to define the label allocated by a device to advertise reachability for the specific anycast prefix segment. The value of this label is derived by applying the SID index associated with the anycast prefix segment as an offset to the SRGB of the specific device. Table 1 below shows the labels allocated by the various devices in Figure 2 for the anycast prefix segment with SID 100.

| Anycast-SID | Device | SRGB | Anycast-Segment-Label |
|---|---|---|---|
| 100 | R1 | 7000-8000 | 7100 |
| 100 | A1 | 1000-2000 | 1100 |
| 100 | A2 | 2000-3000 | 2100 |
| 100 | A3 | 3000-4000 | 3100 |
| 100 | A4 | 4000-5000 | 4100 |
| 100 | R3 | 6000-7000 | 6100 |

Table 1: Anycast Segment Label Allocation

### 3.2. Virtual SID Label Lookup Table

When a MPLS packet on the wire first hits a device, the forwarding hardware reads the topmost label in the MPSL header and looks up the default label lookup table associated with the interface on which the label has been received. This table is generally called LFIB. The range of labels found in the LFIB constitutes the default label space.

This document introduces a separate virtual label lookup table (hereafter referred to as Virtual LFIB or V-LFIB), that represents a label space which is also separate from the actual label space represented by the default LFIB. The label value may be present in both the default and Virtual LFIB. However the forwarding semantics associated with the label under the default and Virtual LFIB may not be same. Following are the fields of a typical entry of this table.

- SID-Index: The SID index assoicated with a prefix segment originated by another device in the same network. This is also the key field for this table.

- Forwarding Semantics: This is once again one or more tuples of following items.

  * Outgoing-Label: The label(s) allocated by the neighbor device(s) on the shortest-path to the topologically nearest originator(s) of the prefix segment.

  * Outgoing-link: The link(s) connecting the device to the neighbor device(s) on the shortest path to the topologically nearest originator(s) of the prefix segment.

This document proposes that, any device, when provisioned with one or more anycast prefix segment (address and SID), it MUST create a Virtual LFIB table. Such a device MUST add an entry in the Virtual LFIB for each unicast and anycast prefix segments learnt from a remote device, if and only if the same prefix has not been provisioned on the device. The device SHOULD NOT add a entry for any of the Anycast or Node prefix segments that it has advertised itself. However if the device has learnt any anycast prefix segment from a remote device, and the same is not provisioned on this device, the device MUST include the same in the Virtual LFIB table.

In cases where a prefix segment is reachable via multiple shortest paths on a given device, the corresponding entry for the prefix SID MUST have as many forwarding entries in the Virtual LFIB table as the number of shortest-paths found for the corresponding prefix on the device. .

Figure 3 below shows how the Virtual LFIB table on each of devices in group A should look like. Please note that some of the prefix segments has multiple forwarding semantics associated with them. For example, on device A1, the prefix SID 10 (originated by PE3) is reachable through its neighbors A3 and A4. And as per the SRGB advertised by A3 and A4, the labels allocated by A3 and A4 are 3030 and 4030 respectively. Hence A1 has added two forwarding entries for the prefix SID 30 in its Virtual LFIB table.

Also please note that none of the devices in the anycast group have included the anycast SID 100 in the Virtual LFIB table, since the same has already been provisioned on these devices.

When a device receives a MPLS packet with the anycast segment label associated with one of the anycast prefix segments provisioned on the same device, the device MUST use the Virtual LFIB table to lookup the next label that follows the anycast segment label in the stack of labels found in the MPLS header. Refer to Section 3.5 for more details.

Following forwarding instructions MUST be installed in the MPLS data-plane for each entry in the Virtual LFIB entry.

- If the label at the top of the stack matches any of the prefix SIDs in the Virtual LFIB table,
  - If there are multiple forwarding tuples associated with matching table entry,
    - Select one forwarding tuple. (Criteria to select one is outside the scope of this document.)
  - Else,
    - Select the single forwarding tuple available.
  - Replace the Prefix SID index found at top of the MPLS label stack in the packet received, with the ’Outgoing-label’ from the selected forwarding tuple.
  - Forward the modified packet onto the ’Outgoing-link’ as specified in the selected forwarding tuple.
  - Ensure the next label lookup is launched on the default LFIB table.

| Device | Prefix SID | Forwarding Semantics | |
|---|---|---|---|
| | | Outgoing-Label | Outgoing-Link |
| A1 | 10 | 7010 | A1->R1 |
| | 20 | 7020 | A1->R1 |
| | 30 | 3030 | A1->A3 |
| | | 4030 | A1->A4 |
| | 40 | 3040 | A1->A3 |
| | | 4040 | A1->A4 |
| A2 | 10 | 7010 | A2->R1 |
| | 20 | 7020 | A2->R1 |
| | 30 | 3030 | A2->A3 |
| | | 4030 | A2->A4 |
| | 40 | 3040 | A2->A3 |
| | | 4040 | A2->A4 |
| A3 | 10 | 1010 | A3->A1 |
| | | 2010 | A3->A2 |
| | 20 | 1020 | A3->A1 |
| | | 2020 | A3->A2 |
| | 30 | 6030 | A3->R3 |
| | 40 | 6040 | A3->R3 |
| A4 | 10 | 1010 | A4->A1 |
| | | 2010 | A4->A2 |
| | 20 | 1020 | A4->A1 |
| | | 2020 | A4->A2 |
| | 30 | 6030 | A4->R3 |
| | 40 | 6040 | A4->R3 |

Figure 3: Virtual LFIB Table Setup

### 3.3. Label Stack Computation

Any MPLS device that tries to encapsulate any kind of traffic into a SPRING-based MPLS payload (hereafter referred to as the ingress device) and steer it through a series of SPRING adjacency and/or unicast/anycast prefix segments, needs to compute an appropriate stack of MPLS labels and put it in the outgoing packet. Alternatively, in a SDN environment, the SDN controller may need to compute the label stack and install it on the ingress device.

However in both cases, as illustrated in Section 2, for a given ingress device (e.g. PE1 or PE2), there maybe multiple topolgically nearest devices in a specific anycast group (e.g. A1 and A2), even through there is only out-going link from the source device(e.g. PE1->R1 or PE2-R1). In such case, when the ingress device (or the SDN controller) wants to steer a packet through the anycast group A, it can use the anycast segment label advertised by the downstream neighbor of the ingress device for the specific anycast prefix segment. Since the packet may reach any one of the multiple devices in the group and each of them may have a separate SRGB label range, choosing the MPLS label for the next segment providing reachability to the final destination. Also, since the packet steered through a anycast segment can reach of any of the member device in the anycast group, it is sufficient to assume that the ingress (or the controller) cannot place an adjacency segment immediately after a anycast segment in the outgoing packet.

This document proposes the ingress device (or the SDN controller) to directly use the SID as the label for a prefix segment (can be another anycast)that immediately follows a given anycast segment already encoded into the label stack of the outgoing MPLS packet. The ingress (or the controller) MUST follow the algorithm below to compute the label-stack it must use to steer a packet through a list of SPRING segments.

- Set ’last_segment’ ==> NONE.
- For [all ’segments’ in Segment_List]
  - If {’segment’.type == Adjacency_Segment}
    - Set ’label’ ==> ’segment’.Adjacency_Segment_Label.
  - Else
    - If {’last_segment’.type == Anycast_Prefix_Segment}
      - Set ’label’ ==> ’segment’.SID_index.

+ Else

- Set 'label' ==> 'Prefix_Segment_Label'.

o Add 'label' to 'label_stack'.

### 3.4. Advertising Anycast Prefix Segments

Like unicast prefix segments, anycast prefix segments SHOULD be advertised in IGP Link-state advertsements using IGP protocol extension for SPRING specified in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions]. This document does not propose any protocol extension for advertising anycast prefix segments.

However when advertising the anycast segments, the originating device MUST set the corresponding P-Flag(No-PHP) in ISIS Prefix-SID SubTLV and/or the NP-Flag (No-PHP) in OSPFv2 and OSPFv3 Prefix-SID SubTLV to 1 and the E-Flag in the same SubTLVs to 0. Please refer to following for more details on usage of these flags.

o ISIS Prefix-SID SubTLV [I-D.ietf-isis-segment-routing-extensions]

o OSPFv2 Prefix-SID SubTLV [I-D.ietf-ospf-segment-routing-extensions]

o OSPFv3 Prefix-SID SubTLV [I-D.ietf-ospf-ospfv3-segment-routing-extensions]

The proposal above, ensures that a MPLS packet sent to (or taking transit through) a given anycast group, always arrives at the topologically nearest device in the group, with a label that is derived from the device's SRGB, and the SID associated with the corresponding anycast prefix segment.

In Figure 2, when PE1 or PE2 intends to steer a packet destined for PE3 or PE4, through the anycast group A (SID 100), it needs to forward the packet to R1 (SRGB:7000-8000), after putting the label 7100 (derived from R1's SRGB), at top of the label stack in the MPLS header. However when the same packet is forwarded to A1 or A2 (topologicaly nearest devices in group A), R1 shall not POP (or remove) the label 7100. Instead R1 shall replace it with the label 1100 (while forwarding to A1) or 2100 (while forwarding to A2).

### 3.5. Programming Anycast Prefix Segments

The proposal specified in Section 3.4, ensures that a MPLS packet destined to (or steered via) a anycast prefix segment always arrives at the nearest device in the anycast group with a label derived from the device's SRGB and the SID associated with the corresponding anycast prefix segment, as the top-most label label stack in its MPLS header. If this label is also the bottom-most label (S=1), it means packet has been destined to the anycast segment, and should be consumed by the local device. If the label is not the bottom-most label (S=0), the packet must be forwarded to the next segment, for which the next label in the stack should be consulted. However Section 3.3 specifies that the next label in such case, shall be directly the SID associated with the next segment. Since the SID associated with a prefix segment may directly collide with another label in the default LFIB table, Section 3.2 also proposed to have a Virtual LFIB table to provide a separate label-space for looking up the next label.

This document specifies that a device provisioned with a given prefix segment index MUST implement following forwarding semantics for the anycast segment label (refer to Section 3.1) associated with the anycast prefix segment.

- If the label at the top the stack is a anycast segment label,
  - Pop the label.
  - If bottom-most label in the stack (S=1),
    - Send it to host stack for local consumption, as usual.
  - Else if not the bottom-most label in the stack (S=0),
    - Set the Virtual LFIB table as the lookup table for the next label lookup.
    - Launch a lookup for the next label in the stack.
- Else
  - Lookup the label in the default LFIB table as usual.

### 3.6. Packet Flow

Figure 4 below ilustrate how SPRING-based MPLS packets destined for PE3 and sourced by PE1 are expected to flow theough when PE1

encapsulates the packet with an appropriate label stack to steer it through group A devices only

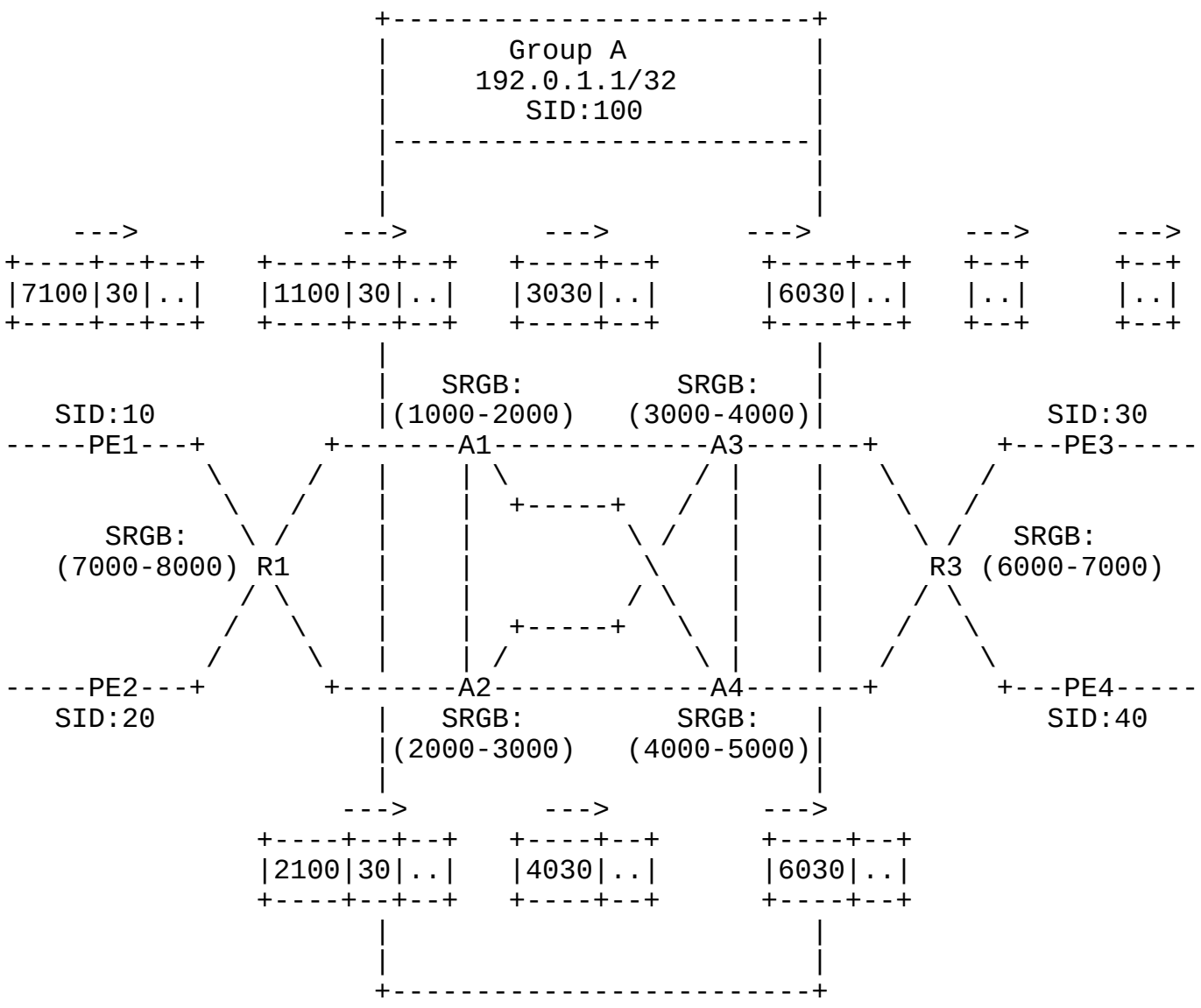


Figure 4: Packet Flow through MPLS-based SPRING Anycast Segments

## 4. Acknowledgements

Many many thanks to Shraddha Hegde for her valuable inputs.

## 5. IANA Considerations

N/A. - No protocol changes are proposed in this document.

## 6. Security Considerations

This document does not introduce any change in any of the protocol specifications. It simply proposes additional inequalities for selecting LFAs for multi-homed prefixes.

## 7. References

### 7.1. Normative References

[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.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-04 (work in progress), May 2015.

[I-D.ietf-ospf-ospfv3-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPFv3 Extensions for Segment Routing", draft-ietf-ospf-ospfv3-segment-routing-extensions-02 (work in progress), February 2015.

[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-04 (work in progress), February 2015.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-03 (work in progress), May 2015.

[I-D.ietf-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Shakir, R., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-01 (work in progress), May 2015.

[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-06 (work in progress), May 2015.

## Authors' Addresses

Pushpasis Sarkar (editor)
Juniper Networks, Inc.
Electra, Exora Business Park
Bangalore, KA  560103
India

Email: psarkar@juniper.net

Hannes Gredler
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA  94089
US

Email: hannes@juniper.net