IDR
Internet-Draft
Intended status: Standards Track
Expires: January 21, 2016

K. Patel
S. Previdi
C. Filsfils
A. Sreekantiah
Cisco Systems
S. Ray
Unaffiliated
H. Gredler
Juniper Networks
July 20, 2015

Segment Routing Prefix SID extensions for BGP
draft-keyupate-idr-bgp-prefix-sid-05

Abstract

Segment Routing (SR) architecture allows a node to steer a packet flow through any topological path and service chain by leveraging source routing. The ingress node prepends a SR header to a packet containing a set of "segments". Each segment represents a topological or a service-based instruction. Per-flow state is maintained only at the ingress node of the SR domain.

This document describes the BGP extension for announcing BGP Prefix Segment Identifier (BGP Prefix SID) information.

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] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without any normative meaning.

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 21, 2016.

Table of Contents

1. Segment Routing Documents

The main references for this document are the SR architecture defined in [I-D.ietf-spring-segment-routing] and the related use case illustrated in [I-D.filsfils-spring-segment-routing-msdc].

The Segment Routing Egress Peer Engineering architecture is described in [I-D.filsfils-spring-segment-routing-central-epe].

The Segment Routing Egress Peer Engineering BGPLS extensions are described in [I-D.ietf-idr-bgpls-segment-routing-epe].

2. Introduction

Segment Routing (SR) architecture leverages the source routing paradigm. A group of inter-connected nodes that use SR forms a SR domain. The ingress node of the SR domain prepends a SR header containing "segments" to an incoming packet. Each segment represents a topological instruction such as "go to prefix P following shortest path" or a service instruction (e.g.: "pass through deep packet inspection"). By inserting the desired sequence of instructions, the ingress node is able to steer a packet via any topological path and/or service chain; per-flow state is maintained only at the ingress node of the SR domain.

Each segment is identified by a Segment Identifier (SID). As described in [I-D.ietf-spring-segment-routing], when SR is applied to the MPLS dataplane the SID consists of a label while when SR is applied to the IPv6 dataplane the SID consists of an IPv6 prefix (see [I-D.previdi-6man-segment-routing-header]).

A BGP-Prefix Segment (aka BGP-Prefix-SID), is a BGP segment attached to a BGP prefix. A BGP-Prefix-SID is always global within the SR/BGP domain and identifies an instruction to forward the packet over the ECMP-aware best-path computed by BGP to the related prefix. The BGP-Prefix-SID is the identifier of the BGP prefix segment.

This document describes the BGP extension to signal the BGP-Prefix-SID. Specifically, this document defines a new BGP attribute known as the BGP Prefix SID attribute and specifies the rules to originate, receive and handle error conditions of the new attribute.

As described in [I-D.filsfils-spring-segment-routing-msdc], the newly proposed BGP Prefix-SID attribute can be attached to prefixes from AFI/SAFI:

Multiprotocol BGP labeled IPv4/IPv6 Unicast ([RFC3107]).

Multiprotocol BGP ([RFC4760]) unlabeled IPv6 Unicast.

[I-D.filsfils-spring-segment-routing-msdc] describes use cases where the Prefix-SID is used for the above AFI/SAFI.

3. BGP-Prefix-SID

The BGP-Prefix-SID attached to a BGP prefix P represents the instruction "go to Prefix P" along its BGP bestpath (potentially ECMP-enabled).

3.1. MPLS Prefix Segment

The BGP Prefix Segment is realized on the MPLS dataplane in the following way:

According to [I-D.ietf-spring-segment-routing], each BGP speaker is configured with a label block called the Segment Routing Global Block (SRGB). The SRGB of a node is a local property and could be different on different speakers.

As described in [I-D.filsfils-spring-segment-routing-msdc] the operator assigns a globally unique "index", L_I, to a locally sourced prefix of a BGP speaker N which is advertised to all other BGP speakers in the SR domain.

The index L_I is a 32 bit offset in the SRGB. Each BGP speaker derives its local MPLS label, L, by adding L_I to the start value of its own SRGB, and programs L in its MPLS dataplane as its incoming/local label for the prefix.

If the BGP speakers are configured with the same SRGB start value, they will all program the same MPLS label for a given prefix P. This has the effect of having a single label for prefix P across all BGP speakers despite that the MPLS paradigm of "local label" is preserved and this clearly simplifies the deployment and operations of traffic engineering in BGP driven networks, as described in [I-D.filsfils-spring-segment-routing-msdc].

If the BGP speakers cannot be configured with the same SRGB, the proposed BGP Prefix-SID attribute allows the advertisement of the SRGB so each node can advertise the SRGB it's configured with. The drawbacks of the use case where BGP speakers have different SRGBs are documented in [I-D.filsfils-spring-segment-routing-msdc].

In order to advertise the label index of a given prefix P and, optionally, the SRGB, a new extension to BGP is needed: the BGP

Prefix SID attribute. This extension is described in subsequent sections.

3.2. IPv6 Prefix Segment

As defined in [I-D.previdi-6man-segment-routing-header], in SR for the IPv6 dataplane, the SRGB consists of the set of IPv6 addresses used within the SR domain (as described in [I-D.previdi-6man-segment-routing-header]). Therefore the BGP speaker willing to process SR IPv6 packets MUST advertise an IPv6 prefix with the attached Prefix SID attribute and related SR IPv6 flag (see subsequent section).

As described in [I-D.filsfils-spring-segment-routing-msdc], when SR is used over an IPv6 dataplane, the BGP Prefix Segment is instantiated by an IPv6 prefix originated by the BGP speaker.

Each node advertises a globally unique IPv6 address representing itself in the domain. This prefix (e.g.: its loopback interface address) is advertised to all other BGP speakers in the SR domain.

Also, each node MUST advertise its support of Segment Routing for IPv6 dataplane. This is realized using the Prefix SID Attribute defined below.

4. BGP-Prefix-SID Attribute

The BGP Prefix SID attribute is an optional, transitive BGP path attribute. The attribute type code is to be assigned by IANA (suggested value: 40). The value field of the BGP-Prefix-SID attribute has the following format:

The value field of the BGP Prefix SID attribute is defined here to be a set of elements encoded as "Type/Length/Value" (i.e., a set of TLVs). Following TLVs are defined:

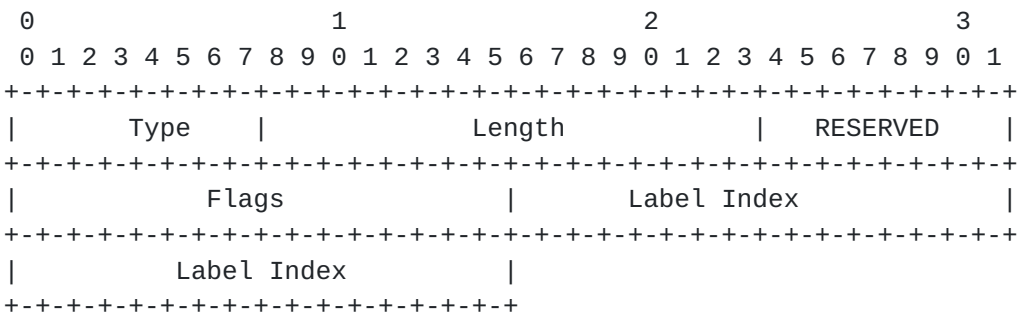
- Label-Index TLV
- IPv6 SID TLV
- Originator SRGB TLV

Label-Index and Originator SRGB TLVs are used only when SR is applied to the MPLS dataplane.

IPv6 SID TLV is used only when SR is applied to the IPv6 dataplane.

4.1. Label-Index TLV

The Label-Index TLV MUST be present in the Prefix-SID attribute attached to Labeled IPv4/IPv6 unicast prefixes ([RFC3107]) and has the following format:

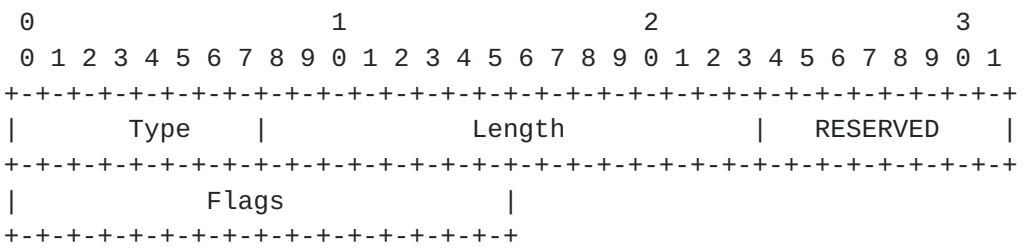


where:

- Type is 1.
- Length: is 7, the total length of the value portion of the TLV.
- RESERVED: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.
- Flags: 16 bits of flags. None are defined at this stage of the document. The flag field SHOULD be clear on transmission and MUST be ignored at reception.
- Label Index: 32 bit value representing the index value in the SRGB space.

4.2. IPv6 SID

The Label-Index TLV MUST be present in the Prefix-SID attribute attached to MP-BGP unlabeled IPv6 unicast prefixes ([RFC4760]) and has the following format:



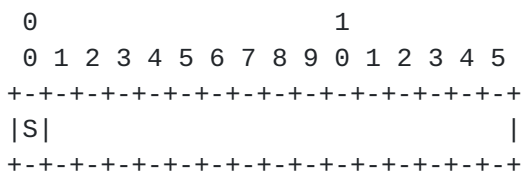
where:

- Type is 2.

- Length: is 3, the total length of the value portion of the TLV.

- RESERVED: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.

- Flags: 16 bits of flags defined as follow:



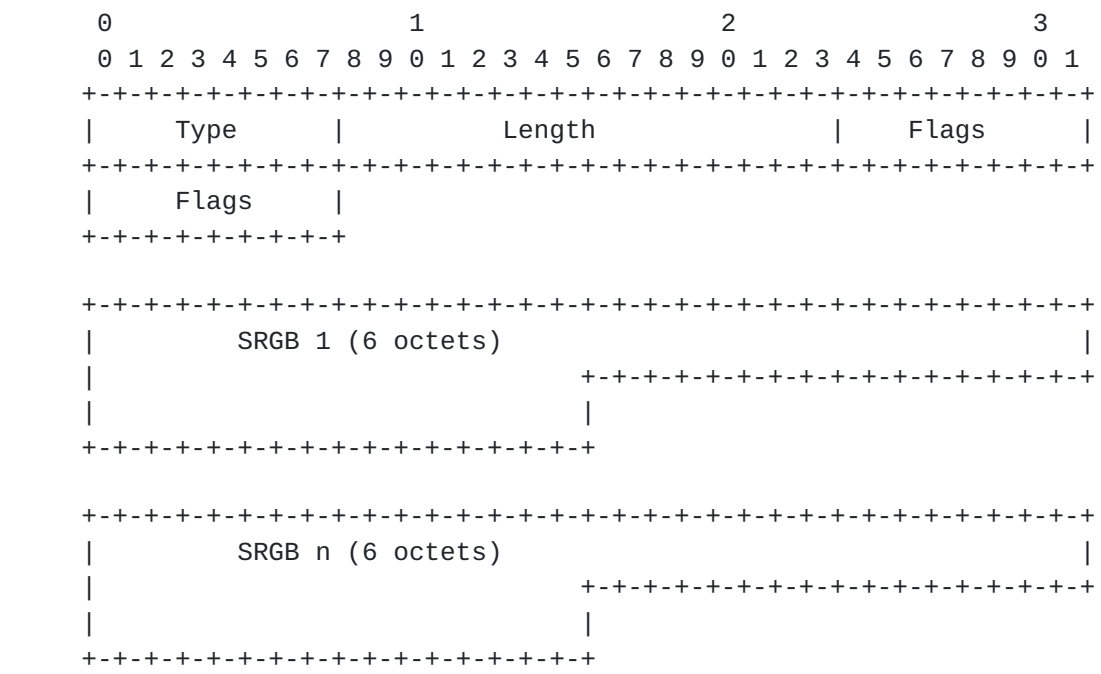
where:

  - S flag: if set then it means that the BGP speaker attaching the Prefix-SID Attribute to a prefix is capable of processing the IPv6 Segment Routing Header (SRH, [I-D.previdi-6man-segment-routing-header]) for the segment corresponding to the originated IPv6 prefix. The use case leveraging the S flag is described in [I-D.filsfils-spring-segment-routing-msdc].

The other bits of the flag field SHOULD be clear on transmission an MUST be ignored at reception.

4.3. Originator SRGB TLV

The Originator SRGB TLV is an optional TLV and has the following format:



where:

- Type is 3.
- Length is the total length of the value portion of the TLV: 2 + multiple of 6.
- Flags: 16 bits of flags. None are defined in this document. Flags SHOULD be clear on transmission an MUST be ignored at reception.
- SRGB: 3 octets of base followed by 3 octets of range. Note that SRGB field MAY appear multiple times.

The Originator SRGB TLV contains the SRGB of the router originating the prefix to which the BGP Prefix SID is attached and MUST be kept in the Prefix-SID Attribute unchanged during the propagation of the BGP update.

The originator SRGB describes the SRGB of the node where the BGP Prefix Segment end. It is used to build SRTE policies when different SRGB's are used in the fabric ([I-D.filsfils-spring-segment-routing-msdc]).

The originator SRGB may only appear on Prefix-SID attribute attached to prefixes of SAFI 4 (labeled unicast, [RFC3107]).

5. Receiving BGP-Prefix-SID Attribute

A BGP speaker receiving a BGP Prefix-SID attribute from an EBGP neighbor residing outside the boundaries of the SR domain, SHOULD discard the attribute unless it is configured to accept the attribute from the EBGP neighbor. A BGP speaker MAY log an error for further analysis when discarding an attribute.

5.1. MPLS Dataplane: Labeled Unicast

The Prefix-SID attribute MUST contain the Label-Index TLV and MAY contain the Originator SRGB. A BGP Prefix-SID attribute received without a Label-Index TLV MUST be considered as "unacceptable" by the receiving speaker.

A BGP speaker may be locally configured with an SRGB=[GB_S, GB_E]. The preferred method for deriving the SRGB is a matter of local router configuration.

Given a label index L_I, we call L = L_I + GB_S as the derived label. A BGP Prefix-SID attribute is called "unacceptable" for a speaker M if the derived label value L lies outside the SRGB configured on M. Otherwise the Label Index attribute is called "acceptable" to speaker M.

The mechanisms through which a given label_index value is assigned to a given prefix are outside the scope of this document. The label-index value associated with a prefix is locally configured at the BGP router originating the prefix.

The Prefix-SID attribute MUST contain the Label-Index TLV and MAY contain the Originator SRGB TLV. A BGP Prefix-SID attribute received without a Label-Index TLV MUST be considered as "unacceptable" by the receiving speaker.

When a BGP speaker receives a path from a neighbor with an acceptable BGP Prefix-SID attribute, it SHOULD program the derived label as the local label for the prefix in its MPLS dataplane. In case of any error, a BGP speaker MUST resort to the error handling rules specified in Section 7. A BGP speaker MAY log an error for further analysis.

When a BGP speaker receives a path from a neighbor with an unacceptable BGP Prefix-SID attribute, for the purpose of label allocation, it SHOULD treat the path as if it came without a Prefix-SID attribute. A BGP speaker MAY choose to assign a local (also called dynamic) label (non-SRGB) for such a prefix. A BGP speaker MAY log an error for further analysis.

A BGP speaker receiving a prefix with a Prefix-SID attribute and a label NLRI field of implicit-null from a neighbor MUST adhere to standard behavior and program its MPLS dataplane to pop the top label when forwarding traffic to the prefix. The label NLRI defines the outbound label that MUST be used by the receiving node. The Label Index gives a hint to the receiving node on which local/incoming label the BGP speaker SHOULD use.

5.2. IPv6 Dataplane

When a SR IPv6 BGP speaker receives a IPv6 Unicast BGP Update with a prefix having the BGP Prefix SID attribute attached, it checks whether the IPv6 SID TLV is present and if the S-flag is set. If the IPv6 SID TLV is not present or if the S-flag is not set, then the Prefix-SID attribute MUST be considered as "unacceptable" by the receiving speaker.

The Originator SRGB MUST be ignored on reception.

A BGP speaker receiving a BGP Prefix-SID attribute from an EBGP neighbor residing outside the boundaries of the SR domain, SHOULD discard the attribute unless it is configured to accept the attribute from the EBGP neighbor. A BGP speaker MAY log an error for further analysis when discarding an attribute.

6. Announcing BGP-Prefix-SID Attribute

The BGP Prefix-SID attribute MAY be attached to labeled BGP prefixes (IPv4/IPv6) [RFC3107]or to IPv6 prefixes [RFC4760]. In order to prevent distribution of the BGP Prefix-SID attribute beyond its intended scope of applicability, attribute filtering MAY be deployed.

6.1. MPLS Dataplane: Labeled Unicast

A BGP speaker that originates a prefix attaches the Prefix-SID attribute when it advertises the prefix to its neighbors. The value of the Label-Index in the Label-Index TLV is determined by configuration.

A BGP speaker that originates a Prefix-SID attribute MAY optionally announce Originator SRGB TLV along with the mandatory Label-Index TLV. The content of the Originator SRGB TLV is determined by the configuration.

Since the Label-index value must be unique within an SR domain, by default an implementation SHOULD NOT advertise the BGP Prefix-SID attribute outside an Autonomous System unless it is explicitly configured to do so.

A BGP speaker that advertises a path received from one of its neighbors SHOULD advertise the Prefix-SID received with the path without modification regardless of whether the Prefix-SID was acceptable. If the path did not come with a Prefix-SID attribute, the speaker MAY attach a Prefix-SID to the path if configured to do so. The content of the TLVs present in the Prefix-SID is determined by the configuration.

In all cases, the label field of the NLRI ([RFC3107], [RFC4364]) MUST be set to the local/incoming label programmed in the MPLS dataplane for the given prefix. If the prefix is associated with one of the BGP speakers interfaces, this label is the usual MPLS label (such as the implicit or explicit NULL label).

6.2. IPv6 Dataplane

A BGP speaker that originates a prefix attaches the Prefix-SID attribute when it advertises the prefix to its neighbors. The IPv6 SID TLV MUST be present and the S-flag MUST be set.

A BGP speaker that advertises a path received from one of its neighbors SHOULD advertise the Prefix-SID received with the path without modification regardless of whether the Prefix-SID was acceptable. If the path did not come with a Prefix-SID attribute, the speaker MAY attach a Prefix-SID to the path if configured to do so. The IPv6-SID TLV MUST be present in the Prefix-SID and with the S-flag set.

7. Error Handling of BGP-Prefix-SID Attribute

When a BGP Speaker receives a BGP Update message containing a malformed BGP Prefix-SID attribute, it MUST ignore the received BGP Prefix-SID attributes and not pass it to other BGP peers. This is equivalent to the -attribute discard- action specified in [I-D.ietf-idr-error-handling]. When discarding an attribute, a BGP speaker MAY log an error for further analysis.

If the BGP Prefix-SID attribute appears more than once in an BGP Update message message, then, according to [I-D.ietf-idr-error-handling], all the occurrences of the attribute other than the first one SHALL be discarded and the BGP Update message shall continue to be processed.

When a BGP speaker receives an unacceptable Prefix-SID attribute, it MAY log an error for further analysis.

8. IANA Considerations

This document defines a new BGP path attribute known as the BGP Prefix-SID attribute. This document requests IANA to assign a new attribute code type (suggested value: 40) for BGP the Prefix-SID attribute from the BGP Path Attributes registry.

This document defines two new TLVs for BGP Prefix-SID attribute. These TLVs need to be registered with IANA. We request IANA to create a new registry for BGP Prefix-SID Attribute TLVs as follows:

Under "Border Gateway Protocol (BGP) Parameters" registry, "BGP Prefix SID attribute Types" Reference: draft-keyupate-idr-bgp-prefix-side-05 Registration Procedure(s): Values 1-254 First Come, First Served, Value 0 and 255 reserved

Value	Type	Reference
0	Reserved	draft-keyupate-idr-bgp-prefix-side-05
1	Label-Index	draft-keyupate-idr-bgp-prefix-side-05
2	IPv6 SID	draft-keyupate-idr-bgp-prefix-side-05
3	Originator SRGB	draft-keyupate-idr-bgp-prefix-side-05
4-254	Unassigned	
255	Reserved	draft-keyupate-idr-bgp-prefix-side-05

9. Security Considerations

This document introduces no new security considerations above and beyond those already specified in [RFC4271] and [RFC3107].

10. Acknowledgements

The authors would like to thanks Satya Mohanty and Acee Lindem for their contribution to this document.

11. Change Log

Initial Version: Sep 21 2014

12. References

12.1. Normative References

[I-D.ietf-idr-error-handling]
Chen, E., Scudder, J., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", draft-ietf-idr-error-handling-19 (work in progress), April 2015.

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

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <http://www.rfc-editor.org/info/rfc3107>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

12.2. Informative References

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

[I-D.filsfils-spring-segment-routing-msdc]
Filsfils, C., Previdi, S., Mitchell, J., Aries, E., Lapukhov, P., Gaya, G., Afanasiev, D., Laberge, T., Nkposong, E., Nanduri, M., Uttaro, J., and S. Ray, "BGP-Prefix Segment in large-scale data centers", draft-filsfils-spring-segment-routing-msdc-02 (work in progress), July 2015.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., Patel, K., Dong, J., and M. Chen, "Segment Routing Egress Peer Engineering BGP-LS Extensions", draft-ietf-idr-bgpls-segment-routing-epe-00 (work in progress), June 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.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.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <http://www.rfc-editor.org/info/rfc4760>.

Authors' Addresses

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: keyupate@cisco.com

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

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfils@cisco.com

Arjun Sreekantiah
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: asreekan@cisco.com

Saikat Ray
Unaffiliated

Email: raysaikat@gmail.com

Hannes Gredler
Juniper Networks

Email: hannes@juniper.net