Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 24, 2016

L. Ginsberg
P. Psenak
S. Previdi
Cisco Systems
M. Pilka

June 22, 2016

Segment Routing Conflict Resolution
draft-ietf-spring-conflict-resolution-01.txt

Abstract

In support of Segment Routing (SR) routing protocols advertise a variety of identifiers used to define the segments which direct forwarding of packets. In cases where the information advertised by a given protocol instance is either internally inconsistent or conflicts with advertisements from another protocol instance a means of achieving consistent forwarding behavior in the network is required. This document defines the policies used to resolve these occurrences.

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 24, 2016.

Table of Contents

1. Introduction

Segment Routing (SR) as defined in [SR-ARCH] utilizes forwarding instructions called "segments" to direct packets through the network. Depending on the forwarding plane architecture in use, routing protocols advertise various identifiers which define the permissible values which can be used as segments, which values are assigned to

specific prefixes, etc. Where segments have global scope it is necessary to have non-conflicting assignments - but given that the advertisements may originate from multiple nodes the possibility exists that advertisements may be received which are either internally inconsistent or conflicting with advertisements originated by other nodes. In such cases it is necessary to have consistent resolution of conflicts network-wide in order to avoid forwarding loops.

The problem to be addressed is protocol independent i.e., segment related advertisements may be originated by multiple nodes using different protocols and yet the conflict resolution MUST be the same on all nodes regardless of the protocol used to transport the advertisements.

The remainder of this document defines conflict resolution policies which meet these requirements. All protocols which support SR MUST adhere to the policies defined in this document.

2. SR Global Block Inconsistency

In support of an MPLS dataplane routing protocols advertise an SR Global Block (SRGB) which defines a set of label ranges reserved for use by the advertising node in support of SR. The details of how protocols advertise this information can be found in the protocol specific drafts e.g., [SR-OSPF], [SR-OSPFv3], and [SR-IS-IS]. However the protocol independent semantics are illustrated by the following example:

The originating router advertises the following ranges:

```
Range 1: (100, 199)
Range 2: (1000, 1099)
Range 3: (500, 599)
```

The receiving routers concatenate the ranges and build the Segment Routing Global Block (SRGB) as follows:

```
SRGB = (100, 199)
       (1000, 1099)
       (500, 599)
```

The indeces span multiple ranges:

```
index=0 means label 100
...
index 99 means label 199
index 100 means label 1000
index 199 means label 1099
...
index 200 means label 500
...
```

Note that the ranges are an ordered set - what labels are mapped to a given index depends on the placement of a given label range in the set of ranges advertised.

For the set of ranges to be usable the ranges MUST be disjoint. Sender behavior is defined in various SR protocol drafts such as [SR-IS-IS] which specify that senders MUST NOT advertise overlapping ranges.

Receivers of SRGB ranges MUST validate the SRGB ranges advertised by other nodes. If the advertised ranges do not conform to the restrictions defined in the respective protocol specification receivers MUST ignore all advertised SRGB ranges from that node. Operationally the node is treated as though it did not advertise any SRGB ranges. [SR-MPLS] defines the procedures for mapping global SIDs to outgoing labels.

Note that utilization of local SIDs (e.g. adjacency SIDs) advertised by a node is not affected by the state of the advertised SRGB.

3. SR-MPLS Segment Identifier Conflicts

In support of an MPLS dataplane Segment identifiers (SIDs) are advertised and associated with a given prefix. SIDs may be advertised in the prefix reachability advertisements originated by a routing protocol (PFX) . SIDs may also be advertised by a Segment Routing Mapping Server (SRMS).

Mapping entries have an explicit context which includes the topology and the SR algorithm. A generalized mapping entry can be represented using the following definitions:

Src- PFX or SRMS
Pi - Initial prefix
Pe - End prefix
L - Prefix length
Lx - Maximum prefix length (32 for IPv4, 128 for IPv6)
Si - Initial SID value
Se - End SID value
R - Range value (See Note 1)
T - Topology
A - Algorithm

A Mapping Entry is then the tuple: (Src, Pi/L, Si, R, T, A)
Pe = (Pi + ((R-1) << (Lx-L))
Se = Si + (R-1)

NOTE 1: The SID advertised in a prefix reachability advertisement always has an implicit range of 1.

Conflicts in SID advertisements may occur as a result of misconfiguration. Conflicts may occur either in the set of advertisements originated by a single node or between advertisements originated by different nodes. Conflicts which occur within the set of advertisements (P-SID and SRMS) originated by a single node SHOULD be prevented by configuration validation on the originating node.

When conflicts occur, it is not possible for routers to know which of the conflicting advertisements is "correct". In order to avoid forwarding loops and/or blackholes, there is a need for all nodes to resolve the conflicts in a consistent manner. This in turn requires that all routers have identical sets of advertisements and that they all use the same selection algorithm. This document defines procedures to achieve these goals.

3.1. Conflict Types

Two types of conflicts may occur - Prefix Conflicts and SID Conflicts. Examples are provided in this section to illustrate these conflict types.

3.1.1. Prefix Conflict

When different SIDs are assigned to the same prefix we have a "prefix conflict". Prefix conflicts are specific to mapping entries sharing the same topology and algorithm.

Example PC1

(PFX, 192.0.2.120/32, 200, 1, 0, 0)
(PFX, 192.0.2.120/32, 30, 1, 0, 0)

The prefix 192.0.2.120/32 has been assigned two different SIDs:
200 by the first advertisement
30 by the second advertisement

Example PC2

(PFX, 2001:DB8::1/128, 400, 1, 2, 0)
(PFX, 2001:DB8::1/128, 50, 1, 2, 0)

The prefix 2001:DB8::1/128 has been assigned two different SIDs:
400 by the first advertisement
50 by the second advertisement

Prefix conflicts may also occur as a result of overlapping prefix ranges.

Example PC3

(SRMS, 192.0.2.1/32, 200, 200, 0, 0)
(SRMS, 192.0.2.121/32, 30, 10, 0, 0)

Prefixes 192.0.2.121/32 - 192.0.2.130/32 are assigned two different SIDs:
320 through 329 by the first advertisement
30 through 39 by the second advertisement

Example PC4
(SRMS, 2001:DB8::1/128, 400, 200, 2, 0)
(SRMS, 2001:DB8::121/128, 50, 10, 2, 0)

Prefixes 2001:DB8::121/128 - 2001:DB8::130/128 are assigned two different SIDs:
420 through 429 by the first advertisement
50 through 59 by the second advertisement

Examples PC3 and PC4 illustrate a complication - only part of the range advertised in the first advertisement is in conflict. It is logically possible to isolate the conflicting portion and try to use the non-conflicting portion(s) at the cost of increased implementation complexity.

A variant of the overlapping prefix range is a case where we have overlapping prefix ranges but no actual SID conflict.

Example PC5

(SRMS, 192.0.2.1/32, 200, 200, 0, 0)
(SRMS, 192.0.2.121/32, 320, 10, 0, 0)

(SRMS, 2001:DB8::1/128, 400, 200, 2, 0)
(SRMS, 2001:DB8::121/128, 520, 10, 2, 0)

Although there is prefix overlap between the two IPv4 entries (and the two IPv6 entries) the same SID is assigned to all of the shared prefixes by the two entries.

Given two mapping entries:

(SRC, P1/L1, S1, R1, T1, A1) and
(SRC, P2/L2, S2, R2, T2, A2)

where P1 <= P2

a prefix conflict exists if all of the following are true:

1)(T1 == T2) && (A1 == A2)
2)P1 <= P2
3)The prefixes are in the same address family.
2)L1 == L2
3)(P1e >= P2) && ((S1 + (P2 - P1)) != S2)

3.1.2. SID Conflict

When the same SID has been assigned to multiple prefixes we have a "SID conflict". SID conflicts are independent of address-family, independent of prefix len, independent of topology, and independent of algorithm. A SID conflict occurs when a mapping entry which has previously been checked to have no prefix conflict assigns one or more SIDs that are assigned by another entry which also has no prefix conflicts.

Example SC1

(PFX, 192.0.2.1/32, 200, 1, 0, 0)
(PFX, 192.0.2.222/32, 200, 1, 0, 0)
SID 200 has been assigned to 192.0.2.1/32 by the
first advertisement.
The second advertisement assigns SID 200 to 192.0.2.222/32.

Example SC2

(PFX, 2001:DB8::1/128, 400, 1, 2, 0)
(PFX, 2001:DB8::222/128, 400, 1, 2, 0)
SID 400 has been assigned to 2001:DB8::1/128 by the
first advertisement.
The second advertisement assigns SID 400 to 2001:DB8::222/128

SID conflicts may also occur as a result of overlapping SID ranges.

Example SC3

(SRMS, 192.0.2.1/32, 200, 200, 0, 0)
(SRMS, 198.51.100.1/32, 300, 10, 0, 0)

SIDs 300 - 309 have been assigned to two different prefixes.
The first advertisement assigns these SIDs
to 192.0.2.101/32 - 192.0.2.110/32.
The second advertisement assigns these SIDs to
198.51.100.1/32 - 198.51.100.10/32.

Example SC4
(SRMS, 2001:DB8::1/128, 400, 200, 2, 0)
(SRMS, 2001:DB8:1::1/128, 500, 10, 2, 0)

SIDs 500 - 509 have been assigned to two different prefixes.
The first advertisement assigns these SIDs to
2001:DB8::101/128 - 2001:DB8::10A/128.
The second advertisement assigns these SIDs to
2001:DB8:1::1/128 - 2001:DB8:1::A/128.

Examples SC3 and SC4 illustrate a complication - only part of the range advertised in the first advertisement is in conflict.

3.2. Processing conflicting entries

Two general approaches can be used to process conflicting entries.

1. Conflicting entries can be ignored

2. A standard preference algorithm can be used to choose which of the conflicting entries will be used

The following sections discuss these two approaches in more detail.

Note: This document does not discuss any implementation details i.e. what type of data structure is used to store the entries (trie, radix tree, etc.) nor what type of keys may be used to perform lookups in the database.

3.2.1. Policy: Ignore conflicting entries

In cases where entries are in conflict none of the conflicting entries are used i.e., the network operates as if the conflicting advertisements were not present.

Implementations are required to identify the conflicting entries and ensure that they are not used.

3.2.2. Policy: Preference Algorithm/Quarantine

For entries which are in conflict properties of the conflicting advertisements are used to determine which of the conflicting entries are used in forwarding and which are "quarantined" and not used. The entire quarantined entry is not used.

This approach requires that conflicting entries first be identified and then evaluated based on a preference rule. Based on which entry is preferred this in turn may impact what other entries are considered in conflict i.e. if A conflicts with B and B conflicts with C - it is possible that A does NOT conflict with C. Hence if as a result of the evaluation of the conflict between A and B, entry B is not used the conflict between B and C will not be detected.

3.2.3. Policy: Preference algorithm/ignore overlap only

A variation of the preference algorithm approach is to quarantine only the portions of the less preferred entry which actually conflicts. The original entry is split into multiple ranges. The ranges which are in conflict are quarantined. The ranges which are not in conflict are used in forwarding. This approach adds complexity as the relationship between the derived sub-ranges of the original mapping entry have to be associated with the original entry - and every time some change to the advertisement database occurs the derived sub-ranges have to be recalculated.

3.2.4. Preference Algorithm

The following algorithm is used to select the preferred mapping entry when a conflict exists. Evaluation is made in the order specified. Prefix conflicts are evaluated first. SID conflicts are then evaluated on the Active entries remaining after Prefix Conflicts have been resolved.

1. PFX source wins over SRMS source
2. Smaller range wins
3. IPv6 entry wins over IPv4 entry
4. Longer prefix length wins
5. Smaller algorithm wins

6. Smaller starting address (considered as an unsigned integer value) wins

7. Smaller starting SID wins

8. If topology IDs are NOT identical both entries MUST be ignored

Using smaller range as the highest priority tie breaker makes advertisements with a range of 1 the most preferred. This has the nice property that a single misconfiguration of an SRMS entry with a large range will not be preferred over a large number of advertisements with smaller ranges.

Since topology identifiers are locally scoped, it is not possible to make a consistent choice network wide when all elements of a mapping entry are identical except for the topology. This is why both entries MUST be ignored in such cases (Rule #8 above). Note that Rule #8 only applies when considering SID conflicts since Prefix conflicts are restricted to a single topology.

3.2.5. Example Behavior - Single Topology/Algorithm

The following mapping entries exist:in the database. For brevity, Topology/Algorithm is omitted and assumed to be (0,0) in all entries.

1. (PFX, 192.0.2.1/32, 100, 1)

2. (PFX, 192.0.2.101/32, 200, 1)

3. (SRMS, 192.0.2.1/32, 400, 255) !Prefix conflict with entries 1 and 2

4. (SRMS, 198.51.100.40/32, 200,1) !SID conflict with entry 2

The table below shows what mapping entries will be used in the forwarding plane (Active) and which ones will not be used (Excluded) under the three candidate policies:

Policy	Active Entries	Excluded Entries
Ignore		(PFX,192.0.2.1/32,100,1) (PFX,192.0.2.101/32,200,1) (SRMS,192.0.2.1/32,400,255) (SRMS,198.51.100.40/32,200,1)
Quarantine	(PFX,192.0.1.1/32,100,1) (PFX,192.0.2.101/32,200,1)	(SRMS,192.0.2.1/32,400,255) (SRMS,198.51.100.40/32,200,1)
Overlap-Only	(PFX,192.0.2.1/32,100,1) (PFX,192.0.2.101/32,200,1) *(SRMS,192.0.2.2/32,401,99) *(SRMS,192.0.2.102/32, 501,153)	(SRMS,198.51.100.40/32,200,1) *(SRMS,192.0.2.1/32,400,1) *(SRMS,192.0.2.101/32,500,1)

* Derived from (SRMS,192.0.2.1/32,400,300)

3.2.6. Example Behavior - Multiple Topologies

When using a preference rule the order in which conflict resolution is applied has an impact on what entries are usable when entries for multiple topologies (or algorithms) are present. The following mapping entries exist in the database:

1. (PFX, 192.0.2.1/32, 100, 1, 0, 0) !Topology 0

2. (PFX, 192.0.2.1/32, 200, 1, 0, 0) !Topology 0, Prefix Conflict with entry #1

3. (PFX, 198.51.100.40/32, 200,1,1,0) ! Topology 1, SID conflict with entry 2

The table below shows what mapping entries will be used in the forwarding plane (Active) and which ones will not be used (Excluded) under the Quarantine Policy based on the order in which conflict resolution is applied.

Order	Active Entries	Excluded Entries
Prefix-Conflict First	(PFX,192.0.2.1/32,100,1,0,0) (PFX,198.51.100.40/32,200,1,1,0)	(PFX,192.0.2.101/32,200,1,0)
SID-Conflict First	(PFX,192.0.2.1/32,100,1,0,0)	(PFX,192.0.2.101/32,200,1,0) (PFX,198.51.100.40/32,200,1,1,0)

This illustrates the advantage of evaluating prefix conflicts within a given topology (or algorithm) before evaluating topology (or algorithm) independent SID conflicts. It insures that entries which will be excluded based on intratopology preference will not prevent a SID assigned in another topology from being considered Active.

3.2.7. Evaluation of Policy Alternatives

The previous sections have defined three alternatives for resolving conflicts - ignore, quarantine, and ignore overlap-only.

The ignore policy impacts the greatest amount of traffic as forwarding to all destinations which have a conflict is affected.

Quarantine allows forwarding for some destinations which have a conflict to be supported.

Ignore overlap-only maximizes the destinations which will be forwarded as all destinations covered by some mapping entry (regardless of range) will be able to use the SID assigned by the winning range. This alternative increases implementation complexity as compared to quarantine. Mapping entries with a range greater than 1 which are in conflict with other mapping entries have to internally be split into 2 or more "derived mapping entries". The derived mapping entries then fall into two categories - those that are in conflict with other mapping entries and those which are NOT in conflict. The former are ignored and the latter are used. Each time the underived mapping database is updated the derived entries have to be recomputed based on the updated database. Internal data structures have to be maintained which maintain the relationship between the advertised mapping entry and the set of derived mapping entries. All nodes in the network have to achieve the same behavior regardless of implementation internals.

There is then a tradeoff between a goal of maximizing traffic delivery and the risks associated with increased implementation complexity.

It is the opinion of the authors that "quarantine" is the best alternative.

3.2.8. Guaranteeing Database Consistency

In order to obtain consistent active entries all nodes in a network MUST have the same mapping entry database. Mapping entries can be obtained from a variety of sources.

- SIDs can be configured locally for prefixes assigned to interfaces on the router itself. Only SIDs which are advertised to protocol peers can be considered as part of the mapping entry database.
- SIDs can be received in prefix reachability advertisements from protocol peers. These advertisements may originate from peers local to the area or be leaked from other areas and/or redistributed from other routing protocols.
- SIDs can be received from SRMS advertisements - these advertisements can originate from routers local to the area or leaked from other areas
- In cases where multiple routing protocols are in use mapping entries advertised by all routing protocols MUST be included.

4. Scope of SR-MPLS SID Conflicts

The previous section defines the types of SID conflicts and procedures to resolve such conflicts when using an MPLS dataplane. The mapping entry database used MUST be populated with entries for destinations for which the associated SID will be used to derive the labels installed in the forwarding plane of routers in the network. This consists of entries associated with intra-domain routes.

There are cases where destinations which are external to the domain are advertised by protocol speakers running within that network - and it is possible that those advertisements have SIDs associated with those destinations. However, if reachability to a destination is topologically outside the forwarding domain of the protocol instance then the SIDs for such destinations will never be installed in the forwarding plane of any router within the domain - so such advertisements cannot create a SID conflict within the domain. Such entries therefore MUST NOT be installed in the database used for intra-domain conflict resolution.

Consider the case of two sites "A and B" associated with a given [RFC4364] VPN. Connectivity between the sites is via a provider backbone. SIDs associated with destinations in Site A will never be installed in the forwarding plane of routers in Site B. Reachability between the sites (assuming SR is being used across the backbone) only requires using a SID associated with a gateway PE. So a destination in Site A MAY use the same SID as a destination in Site B without introducing any conflict in the forwarding plane of routers in Site A.

Such cases are handled by insuring that the mapping entries in the database used by the procedures defined in the previous section only include entries associated with advertisements within the site.

5. Security Considerations

TBD

6. IANA Consideration

This document has no actions for IANA.

7. Acknowledgements

The authors would like to thank Jeff Tantsura, Wim Henderickx, and Bruno Decraene for their careful review and content suggestions.

8. References

8.1. Normative References

[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>.

[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>.

[SR-IS-IS] "IS-IS Extensions for Segment Routing, draft-ietf-isis-segment-routing-extensions-07(work in progress)", June 2016.

[SR-MPLS] "Segment Routing with MPLS dataplane, draft-ietf-spring-segment-routing-mpls-04(work in progress)", March 2016.

[SR-OSPF] "OSPF Extensions for Segment Routing, draft-ietf-ospf-segment-routing-extensions-08(work in progress)", May 2016.

[SR-OSPFv3] "OSPFv3 Extensions for Segment Routing, draft-ietf-ospf-ospfv3-segment-routing-extensions-05(work in progress)", March 2016.

8.2. Informational References

[SR-ARCH] "Segment Routing Architecture, draft-ietf-spring-segment-routing-08(work in progress)", May 2016.

Authors' Addresses

Les Ginsberg
Cisco Systems
510 McCarthy Blvd.
Milpitas, CA 95035
USA

Email: ginsberg@cisco.com

Peter Psenak
Cisco Systems
Apollo Business Center Mlynske nivy 43
Bratislava 821 09
Slovakia

Email: ppsenak@cisco.com

Stefano Previdi
Cisco Systems
Via Del Serafico 200
Rome 0144
Italy

Email: sprevidi@cisco.com

Martin Pilka

Email: martin@infobox.sk