Workgroup: Network Working Group
Internet-Draft:
draft-ietf-spring-sr-replication-segment-13
Published: 2 March 2023
Intended Status: Standards Track
Expires: 3 September 2023
Authors: D. Voyer, Ed. C. Filsfils
Bell Canada Cisco Systems, Inc.
R. Parekh H. Bidgoli Z. Zhang
Cisco Systems, Inc. Nokia Juniper Networks

SR Replication Segment for Multi-point Service Delivery

Abstract

This document describes the SR Replication segment for Multi-point service delivery. A SR Replication segment allows a packet to be replicated from a Replication Node to Downstream nodes.

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 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 https://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 3 September 2023.

Copyright Notice

Table of Contents

1. Introduction

Replication segment is a new type of segment for Segment Routing [RFC8402], which allows a node (henceforth called a Replication Node) to replicate packets to a set of other nodes (called Downstream Nodes) in a Segment Routing Domain. Replication segments provide building blocks for Point-to-Multipoint Service delivery via SR Point-to-Multipoint (SR P2MP) policy. A Replication segment can replicate packets to directly connected nodes or to downstream nodes (without need for state on the transit routers). This document focuses on the Replication segment building block. The use of one or more stitched Replication segments constructed for SR P2MP Policy tree is specified in [I-D.ietf-pim-sr-p2mp-policy].

1.1. Terminology

*Multi-point Service: A service that has multiple endpoints. A packet is delivered to all the endpoints.

*Replication Segment: A segment in SR domain that replicates packets.

*Replication Node: A node in SR domain which replication packets based on Replication Segment

*Downstream Nodes: A Replication Node replicates packets to a set of Downstream Nodes

*Replication-ID: Identifier of a Replication Segment at Replication Node

*Replication State: This is state of Replication Segment at a Replication Node. It is conceptually a list of replication branches to Downstream nodes. The list can be empty.

*Replication SID: Data plane identifier of a Replication Segment. This is a SR-MPLS label or SRv6 SID.

2. Replication Segment

In a Segment Routing Domain, a Replication segment is a logical construct which connects a Replication Node to a set of Downstream Nodes. A Replication segment is a local segment instantiated at a Replication node. It can be either provisioned locally on a node or programmed by a PCE. Replication segments apply equally to both Segment Routing over MPLS (SR-MPLS) and IPv6 (SRv6).

A Replication segment is identified by the tuple <Replication-ID, Node-ID>, where:

*Replication-ID: An identifier for a Replication segment that is unique in context of the Replication Node.

*Node-ID: The address of the Replication Node that the Replication segment is for. Note that the root of a multi-point service is also a Replication Node.

Replication-ID is a variable length field. In simplest case, it can be a 32-bit number, but it can be extended or modified as required based on specific use of a Replication segment. When the PCE signals a Replication segment to its node, the <Replication-ID, Node-ID> tuple identifies the segment. Examples of such signaling and extension are described in [I-D.ietf-pim-sr-p2mp-policy].

A Replication segment includes the following elements:

- *Replication SID: The Segment Identifier of a Replication segment. This is a SR-MPLS label or a SRv6 SID [RFC8402].

- *Downstream Nodes: Set of nodes in Segment Routing domain to which a packet is replicated by the Replication segment.

- *Replication State: See below.

The Downstream Nodes and Replication State of a Replication segment can change over time, depending on the network state and leaf nodes of a multi-point service that the segment is part of.

Replication SID identifies the Replication segment in the forwarding plane. At a Replication node, the Replication SID operates on local state of Replication segment and the resulting behavior MAY be similar to a Binding SID [RFC9256] of a Segment Routing Policy.

Replication State is a list of replication branches to the Downstream Nodes. In this document, each branch is abstracted to a <Downstream Node, Downstream Replication SID> tuple. <Downstream Node> represents the reachability from the Replication Node to the Downstream Node. In its simplest form, this MAY be specified as an interface or next-hop if downstream node is adjacent to the Replication Node. The reachability may be specified in terms of Flex-Algo path (including the default algo) [RFC9350], or specified by an SR explicit path represented either by a SID-list (of one or more SIDs) or by a Segment Routing Policy [RFC9256]. Downstream Replication SID is the Replication SID of the Replication Segment at the Downstream Node.

A packet is steered into a Replication segment at a Replication Node in two ways:

- *When the Active Segment [RFC8402] is a locally instantiated Replication SID

- *By the root of a multi-point service based on local configuration outside the scope of this document.

In either case, the packet is replicated to each Downstream node in the associated Replication state.

If a Downstream Node is an egress (aka leaf) of the multi-point service, i.e. no further replication is needed, then that leaf node's Replication segment will not have any Replication State i.e. the list of Replication branches is empty. The Replication segment will have an indicator role of the node is Leaf. The operation performed on incoming Replication SID is NEXT. At an egress node,

the Replication SID MAY be used to identify that portion of the multi-point service. Notice that the segment on the leaf node is still referred to as a Replication segment for the purpose of generalization.

A node can be a bud node, i.e. it is a Replication Node and a leaf node of a multi-point service at the same time [I-D.ietf-pim-sr-p2mp-policy]. Replication Segment of a Bud Node has a list of Replication Branches as well as Leaf role indicator.

In principle it is possible for different Replication Segments to replicate packets to the same Replication Segment on a Downstream Node. However, such usage is intentionally left out of scope of this document.

2.1. SR-MPLS data plane

When the Active Segment is a Replication SID, the processing results in a POP operation and lookup of the associated Replication state. For each replication in the Replication state, the operation is a PUSH of the downstream Replication SID and an optional segment list on to the packet to steer the packet to the Downstream node.

For Leaf/Bud nodes local delivery off tree is per local configuration. For some usages, this may involve looking at the next SID for example to get the necessary context.

When the root of a multi-point service steers a packet to a Replication segment, it results in a replication to each Downstream node in the associated replication state. The operation is a PUSH of the replication SID and an optional segment list on to the packet which is forwarded to the downstream node.

SIDs MAY be added before the downstream SR-MPLS Replication SID in order to guide a packet from a non-adjacent SR node to a Replication Node. A Replication Node MAY replicate a packet to a non-adjacent Downstream Node using SIDs it inserts in the copy preceding the downstream Replication SID. The Downstream Node may be leaf node of the Replication Segment, or another Replication Node, or both in case of bud node. A Replication Node MAY use an Anycast SID or BGP PeerSet SID in segment list to send a replicated packet to one downstream Replication node in an Anycast set if and only if all nodes in the set have an identical Replication SID and reach the same set of receivers.. For some use cases, there MAY be SIDs after the Replication SID in the segment list of a packet. These SIDs are used only by the Leaf/Bud nodes to forward a packet off the tree independent of the Replication SID. Coordination regarding the absence or presence and value of context information for Leaf/Bud nodes is outside the scope of this document.

2.2. SRv6 data plane

In SRv6 [RFC8986], the “Endpoint with replication” behavior (End.Replicate for short) replicates a packet and forwards the packet according to a Replication state.

When processing a packet destined to a local Replication SID, the packet is replicated according to the associated replication state to Downstream nodes and/or locally delivered off tree when this is a Leaf/Bud node. IPv6 Hop Limit MUST be decremented and MUST be non-zero to replicate an incoming packet. For replication, the outer header is re-used, and the Downstream Replication SID is written into the outer IPv6 header destination address. If required, an optional segment list may be used on some branches using H.Encaps.Red (while some other branches may not need that). Note that this H.Encaps.Red is independent from the replication segment – it is just used to steer the replicated packet on a traffic engineered path to a Downstream node. The pen-ultimate segment in encapsulating IPv6 header will execute USD flavor of End/End.X behavior and forward the inner (replicated) packet to the Downstream node.

The above also applies when the Replication segment is for the Root node, whose upstream node has placed the Replication-SID in the header. A local application (e.g. MVPN/EVPN) may also apply H.Encaps.Red and then steer the resulting traffic into the segment. Again note that the H.Encaps.Red is independent of the Replication segment – it is the action of the application (e.g. MVPN/EVPN service). If the service is on a Root node, the two H.Encaps mentioned, one for the service and other in the previous paragraph for replication to Downstream node SHOULD be combined for optimization (to avoid extra IPv6 encapsulation).

For Leaf/Bud nodes local delivery off the tree is per Replication SID or next SID (if present in SRH). For some usages, this may involve getting the necessary context either from the next SID (e.g., MVPN with shared tree) or from the replication SID itself (e.g., MVPN with non-shared tree). In both cases, the context association is achieved with signaling and is out of scope of this document

There MAY be SIDs preceding the SRv6 Replication SID in order to guide a packet from a non-adjacent SR node to a Replication Node via an explicit path. A Replication Node MAY steer a replicated packet on an explicit path to a non-adjacent Downstream Node using SIDs it inserts in the copy preceding the downstream Replication SID. The Downstream Node may be leaf node of the Replication Segment, or another Replication Node, or both in case of bud node. For SRv6, as described in above paragraphs, the insertion of SIDs prior to

Replication SID entails a new IPv6 encapsulation with SRH, but this can be optimized on Root node or for compressed SRv6 SIDs. Note that locator of Replication SID is sufficient to guide a packet on IGP shortest path, for default or Flex algo, between non-adjacent nodes. A Replication Node MAY use an Anycast SID or BGP PeerSet SID in segment list to send a replicated packet to one downstream Replication node in an Anycast set if and only if all nodes in the set have an identical Replication SID and reach the same set of receivers. There MAY be SIDs after the Replication SID in the SRH of a packet. These SIDs are used to provide additional context for processing a packet locally at the node where the Replication SID is the Active Segment. Coordination regarding the absence or presence and value of context information for Leaf/Bud nodes is outside the scope of this document.

2.2.1. End.Replicate: Replicate and/or Decapsulate

The "Endpoint with replication and/or decapsulate behavior (End.Replicate for short) is variant of End behavior.

A Replication State conceptually contains following elements:

```
Replication State:
{
  Node-Role: {Head, Transit, Leaf, Bud};
  # On Leaf, replication list is zero length
  Replication-List:
  {
    Downstream Node: <Node-Identifier>;
    Downstream Replication SID: R-SID;
    # Segment-List maybe be empty
    Segment-List: [SID-1, .... SID-N];
  }
}
```

Below is the Replicate function on a packet for Replication State (RS).

```
S01. Replicate(RS, packet)
S02. {
S03.    For each Replication R in RS.Replication-List {
S04.       Make a copy of the packet
S05.       Set IPv6 DA = RS.R-SID
S06.       If RS.Segment-List is not empty {
S07.         # Head node MAY optimize below encap and
S08.         # the encap of packet in a single encap
S09.         Execute H.Encaps or H.Encaps.Red with RS.Segment-List
             on packet copy #RFC 8986 Section 5.1, 5.2
S10.       }
S11.       Submit the packet to the egress IPv6 FIB lookup and
           transmission to the new destination
S12.   }
S13. }
```

Notes:

*The IPv6 destination address in the copy of a packet is set from local state and not from SRH

When N receives a packet whose IPv6 DA is S and S is a local End.Replicate SID, N does:

```
S01.   Lookup FUNCT portion of S to get Replication State RS
S02.   If (IPv6 Hop Limit <= 1) {
S03.     Discard the packet
S04.     # ICMP Time Exceeded is not permitted (Section 2.2.3 below)
S05.   }
S06.   If RS is not found {
S07.     Discard the packet
S08.   }
S09.   Decrement IPv6 Hop Limit by 1
S10.   If (IPv6 NH == SRH and SRH TLVs present) {
S11.     Process SRH TLVs if allowed by local configuration
S12.   }
S13.   Call Replicate(RS, packet)
S14.   If (RS.Node-Role == Leaf or RS.Node-Role == Bud) {
S15.     If (IPv6 NH == SRH and Segments Left > 0) {
S16.       Derive packet processing context(PPC)
           from Segment List[Segments Left - 1]
S17.     } Else {
S18.       Derive packet processing context(PPC)
           from FUNCT of Replication SID
S19.     }
S20.     Remove the outer IPv6 header with all its extension headers
S21.     Process the packet in context of PPC
S21.   }
```

Notes:

*The behavior above MAY result in a packet with partially processed segment list in SRH under some circumstances. Fox example a head node may encode a context SID in a SRH. As per psuedo-code above, a Replication node that receives a packet with local Replication SID will not process the SRH segment list and just forward a copy with unmodified SRH to downstream nodes.

*The packet processing context usually is a FIB table T

Processing the Replication SID may modify, if configured to process TLVs, the "variable-length data" of TLV types that change en route. Therefore, TLVs that change en route are mutable. The remainder of the SRH (Segments Left, Flags, Tag, Segment List, and TLVs that do not change en route) are immutable while processing this SID.

2.2.1.1. HMAC SRH TLV

If a Head Node encodes a context SID in SRH with an optional HMAC SRH TLV [RFC8754], it MUST set the 'D' bit as defined in Section 2.1.2 because the Replication SID is not part of the segment list in SRH.

HMAC generation and verification is as specified in Section 2.1.2.1 of RFC 8754. Verification of HMAC TLV is determined by local configuration. If verification fails, an implementation of Replication SID MUST NOT originate an ICMPv6 error message (parameter problem, code 0). The failure SHOULD be logged (rate limited) and the packet SHOULD be discarded.

2.2.2. OAM Operations

RFC 9259 [RFC9259] specifies procedures for OAM operations like ping and traceroute on SRv6 SIDs.

It is possible to ping a Replication SID of a Leaf/Bud node, assuming the source node knows the Replication SID apriori, directly by putting it in the IPv6 destination address without a SRH or in a SRH as the last segment. While it is not possible to ping a Replication SID of a transit node because transit nodes do not process upper layer headers, it is still possible to ping a Replication SID of Leaf/Bud node of a tree via the Replication SID of intermediate transit nodes. The source of ping MUST compute the ICMPv6 Echo Request checksum using the Replication SID of Leaf/Bud as destination address. The source can then send the Echo Request packet to a transit node's Replication SID. The transit nodes replicate the packet by replacing the IPv6 destination address till the packet reaches the Leaf/Bud node which responds with an ICMPv6

Echo Reply. Appendix A.2.1 illustrates examples of ping to a Replication SID.

Traceroute to a Leaf/Bud Replication SID is not possible due to restriction prohibiting origination of ICMPv6 Time Exceeded error message for a Replication SID as described in the section below.

2.2.3. ICMPv6 Error Messages

ICMPv6 RFC [RFC4443] Section 2.4 states an ICMPv6 error message MUST NOT be originated as a result of receiving a packet destined to an IPv6 multicast address. This is to prevent a storm of ICMPv6 error messages resulting from replicated IPv6 packets from overwhelming a source node. There are two exceptions (1) the Packet Too Big message for Path MTU discovery, and (2) Parameter Problem Message, Code 2 reporting an unrecognized IPv6 option.

An implementation of Replication Segment for SRv6 MUST enforce this same restrictions and exceptions, though this specification does not use any extension header a future extension may do so and MUST support the exception (2) above.

3. Use Cases

In the simplest use case, a single Replication segment includes the root node of a multi-point service and the egress/leaf nodes of the service as all the Downstream Nodes. This achieves Ingress Replication [RFC7988] that has been widely used for MVPN [RFC6513] and EVPN [RFC7432] BUM (Broadcast, Unknown and Multicast) traffic.

Replication segments can also be used as building blocks for replication trees when Replication segments on the root, intermediate Replication Nodes and leaf nodes are stitched together to achieve efficient replication. That is specified in [I-D.ietf-pim-sr-p2mp-policy].

4. Implementation Status

Note to the RFC Editor: Please remove this section and reference to RFC 7942 before publication.

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in RFC 7942 [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors.

This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist. According to RFC 7942 [RFC7942], "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

There are two known implementations of this draft by Cisco and Nokia. Interoperability reports for the implementations are not applicable since this draft does not specify any inter-operable elements of Replication segments.

4.1. Cisco implementation

Cisco Implementation uses Replication Segments defined in this draft as a basis for PCE to compute and establish P2MP trees in SR domain to provide multi-point services. The implementation, based on latest version of this draft, is in production and supports all MUST and SHOULD clauses for SR-MPLS Replication segments. The documentation is available at Cisco documentation and the point of contact is Rishabh Parekh (riparekh@cisco.com).

4.2. Nokia implementation

Nokia has implemented replication SID as defined in this draft to establish P2MP tree in segment routing domain. The implementation supports SR-MPLS encapsulation and has all the Must and SHOULD clause in this draft. The implementation is at general availability maturity and is compliant with the latest version of the draft. The documentation for implementation can be found at Nokia help and the point of contact is hooman.bidgoli@nokia.com.

5. IANA Considerations

IANA has assigned the following codepoint in "SRv6 Endpoint Behaviors" sub-registry of "Segment Routing Parameters" top-level registry for End.Replicate behavior.

Value	Hex	Endpoint behavior	Reference
75	0x004B	End.Replicate	[This.ID]

Table 1: IETF - SRv6 Endpoint Behaviors

6. Security Considerations

The security considerations described in RFC 8402, RFC 8986 and RFC 8754 also apply to this document.

ICMPv6 specification [RFC4443] Section 5.2 describes how the Parameter Problem Message, code 2 exception for ICMPv6 Error message originated for IPv6 multicast destination can be used by a malicious node to cause a denial-of-service attack. Although this specification does not use any extension headers, any future extension doing so is susceptible to the same security consideration.

7. Acknowledgements

The authors would like to acknowledge Siva Sivabalan, Mike Koldychev, Vishnu Pavan Beeram, Alexander Vainshtein, Bruno Decraene, Thierry Couture, Joel Halpern and Ketan Talaulikar for their valuable inputs.

8. Contributors

Clayton Hassen Bell Canada Vancouver Canada

Email: clayton.hassen@bell.ca

Kurtis Gillis Bell Canada Halifax Canada

Email: kurtis.gillis@bell.ca

Arvind Venkateswaran Cisco Systems, Inc. San Jose US

Email: arvvenka@cisco.com

Zafar Ali Cisco Systems, Inc. US

Email: zali@cisco.com

Swadesh Agrawal Cisco Systems, Inc. San Jose US

Email: swaagraw@cisco.com

Jayant Kotalwar Nokia Mountain View US

Email: jayant.kotalwar@nokia.com

Tanmoy Kundu Nokia Mountain View US

Email: tanmoy.kundu@nokia.com

Andrew Stone Nokia Ottawa Canada

Email: andrew.stone@nokia.com

Tarek Saad Cisco Systems Inc. Canada

Email:tsaad@cisco.com

Kamran Raza Cisco Systems, Inc. Canada

Email:skraza@cisco.com

9. References

9.1. 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, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <https://www.rfc-editor.org/info/rfc9256>.

[RFC9259] Ali, Z., Filsfils, C., Matsushima, S., Voyer, D., and M. Chen, "Operations, Administration, and Maintenance (OAM) in Segment Routing over IPv6 (SRv6)", RFC 9259, DOI

10.17487/RFC9259, June 2022, <https://www.rfc-editor.org/info/rfc9259>.

9.2. Informative References

[I-D.filsfils-spring-srv6-net-pgm-illustration]
Filsfils, C., Camarillo, P., Li, Z., Matsushima, S., Decraene, B., Steinberg, D., Lebrun, D., Raszuk, R., and J. Leddy, "Illustrations for SRv6 Network Programming", Work in Progress, Internet-Draft, draft-filsfils-spring-srv6-net-pgm-illustration-04, 30 March 2021, <https://datatracker.ietf.org/doc/html/draft-filsfils-spring-srv6-net-pgm-illustration-04>.

[I-D.ietf-pim-sr-p2mp-policy] Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. J. Zhang, "Segment Routing Point-to-Multipoint Policy", Work in Progress, Internet-Draft, draft-ietf-pim-sr-p2mp-policy-05, 2 July 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-pim-sr-p2mp-policy-05>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

[RFC7988] Rosen, E., Ed., Subramanian, K., and Z. Zhang, "Ingress Replication Tunnels in Multicast VPN", RFC 7988, DOI 10.17487/RFC7988, October 2016, <https://www.rfc-editor.org/info/rfc7988>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

Appendix A. Illustration of a Replication Segment

This section illustrates an example of a single Replication segment. Examples showing Replication segment stitched together to form P2MP tree (based on SR P2MP policy) are in [I-D.ietf-pim-sr-p2mp-policy].

Consider the following topology:

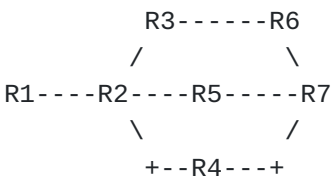

Figure 1: Topology for illustration of Replication Segment

A.1. SR-MPLS

In this example, the Node-SID of a node Rn is N-SIDn and Adjacency-SID from node Rm to node Rn is A-SIDmn. Interface between Rm and Rn is Lmn. The state representation uses "R-SID->Lmn" to represent a packet replication with outgoing replication SID R-SID sent on interface Lmn.

Assume a Replication segment identified with R-ID at Replication Node R1 and downstream Nodes R2, R6 and R7. The Replication SID at node n is R-SIDn. A packet replicated from R1 to R7 has to traverse R4.

The Replication segment state at nodes R1, R2, R6 and R7 is shown below. Note nodes R3, R4 and R5 do not have state for the Replication segment.

Replication segment at R1:

```
Replication segment <R-ID,R1>:
 Replication SID: R-SID1
 Replication State:
   R2: <R-SID2->L12>
   R6: <N-SID6, R-SID6>
   R7: <N-SID4, A-SID47, R-SID7>
```

Replication to R2 steers packet directly to R2 on interface L12. Replication to R6, using N-SID6, steers packet via IGP shortest path to that node. Replication to R7 is steered via R4, using N-SID4 and then adjacency SID A-sID47 to R7.

Replication segment at R2:

```
Replication segment <R-ID,R2>:
 Replication SID: R-SID2
 Replication State:
   R2: <Leaf>
```

Replication segment at R6:

```
Replication segment <R-ID,R6>:
 Replication SID: R-SID6
 Replication State:
   R6: <Leaf>

   Replication segment at R7:

Replication segment <R-ID,R7>:
 Replication SID: R-SID7
 Replication State:
   R7: <Leaf>
```

When a packet is steered into the Replication segment at R1:

- Since R1 is directly connected to R2, R1 performs PUSH operation with just <R-SID2> label for the replicated copy and sends it to R2 on interface L12. R2, as Leaf, performs NEXT operation, pops R-SID2 label and delivers the payload.

- R1 performs PUSH operation with <N-SID6, R-SID6> label stack for the replicated copy to R6 and sends it to R2, the nexthop on IGP shortest path to R6. R2 performs CONTINUE operation on N-SID6 and forwards it to R3. R3 is the penultimate hop for N-SID6; it performs penultimate hop popping, which corresponds to the NEXT operation and the packet is then sent to R6 with <R-SID6> in the label stack. R6, as Leaf, performs NEXT operation, pops R-SID6 label and delivers the payload.

- R1 performs PUSH operation with <N-SID4, A-SID47, R-SID7> label stack for the replicated copy to R7 and sends it to R2, the nexthop on IGP shortest path to R4. R2 is the penultimate hop for N-SID4; it performs penultimate hop popping, which corresponds to the NEXT operation and the packet is then sent to R4 with <A-SID47, R-SID1> in the label stack. R4 performs NEXT operation, pops A-SID47, and delivers packet to R7 with <R-SID7> in the label stack. R7, as Leaf, performs NEXT operation, pops R-SID7 label and delivers the payload.

A.2. SRv6

For SRv6 , we use SID allocation scheme, reproduced below, from Illustrations for SRv6 Network Programming [I-D.filsfils-spring-srv6-net-pgm-illustration]

- 2001:db8::/32 is an IPv6 block allocated by a Regional Internet Registry (RIR) to the operator

- 2001:db8:0::/48 is dedicated to the internal address space

- 2001:db8:cccc::/48 is dedicated to the internal SRv6 SID space

*We assume a location expressed in 64 bits and a function expressed in 16 bits

*Node k has a classic IPv6 loopback address 2001:db8::k/128 which is advertised in the IGP

*Node k has 2001:db8:cccc:k::/64 for its local SID space. Its SIDs will be explicitly assigned from that block

*Node k advertises 2001:db8:cccc:k::/64 in its IGP

*Function :1:: (function 1, for short) represents the End function with PSP support

*Function :Cn:: (function Cn, for short) represents the End.X function from to Node n

Each node k has:

*An explicit SID instantiation 2001:db8:cccc:k:1::/128 bound to an End function with additional support for PSP

*An explicit SID instantiation 2001:db8:cccc:k:Cj::/128 bound to an End.X function to neighbor J with additional support for PSP

*An explicit SID instantiation 2001:db8:cccc:k:Fk::/128 bound to an End.Replicate function

Assume a Replication segment identified with R-ID at Replication Node R1 and downstream Nodes R2, R6 and R7. The Replication SID at node k, bound to an End.Replicate function, is 2001:db8:cccc:k:Fk::/128. A packet replicated from R1 to R7 has to traverse R4.

The Replication segment state at nodes R1, R2, R6 and R7 is shown below. Note nodes R3, R4 and R5 do not have state for the Replication segment. The state representation uses "R-SID->Lmn" to represent a packet replication with outgoing replication SID R-SID sent on interface Lmn.

Replication segment at R1:

```
Replication segment <R-ID,R1>:
 Replication SID: 2001:db8:cccc:1:F1::0
 Replication State:
   R2: <2001:db8:cccc:2:F2::0->L12>
   R6: <2001:db8:cccc:6:F6::0>
   R7: <2001:db8:cccc:4:C7::0, 2001:db8:cccc:7:F7::0>
```

Replication to R2 steers packet directly to R2 on interface L12. Replication to R6, using 2001:db8:cccc:6:F6::0, steers packet via

IGP shortest path to that node. Replication to R7 is steered via R4, using End.X SID 2001:db8:cccc:4:C7::0 at R4 to R7.

Replication segment at R2:

```
Replication segment <R-ID,R2>:
 Replication SID: 2001:db8:cccc:2:F2::0
 Replication State:
   R2: <Leaf>
```

Replication segment at R6:

```
Replication segment <R-ID,R6>:
 Replication SID: 2001:db8:cccc:6:F6::0
 Replication State:
   R6: <Leaf>
```

Replication segment at R7:

```
Replication segment <R-ID,R7>:
 Replication SID: 2001:db8:cccc:7:F7::0
 Replication State:
   R7: <Leaf>
```

When a packet, (A,B2), is steered into the Replication segment at R1:

- Since R1 is directly connected to R2, R1 creates encapsulated replicated copy (2001:db8::1, 2001:db8:cccc:2:F2::0) (A, B2), and sends it to R2 on interface L12. R2, as Leaf, removes outer IPv6 header and delivers the payload.

- R1 creates encapsulated replicated copy (2001:db8::1, 2001:db8:cccc:6:F6::0) (A, B2) then forwards the resulting packet on the shortest path to 2001:db8:cccc:6::/64. R2 and R3 forward the packet using 2001:db8:cccc:6::/64. R6, as Leaf, removes outer IPv6 header and delivers the payload.

- R1 creates encapsulated replicated copy (2001:db8::1, 2001:db8:cccc:4:C7::0) (2001:db8:cccc:7:F7::0; SL=1) (A, B2) and sends it to R2, the nexthop on IGP shortest path to 2001:db8:cccc:4::/64. R2 forwards packet to R4 using 2001:db8:cccc:4::/64. R4 executes End.X function on 2001:db8:cccc:4:C7::0, performs PSP action, removes SRH and sends resulting packet (2001:db8::1, 2001:db8:cccc:7:F7::0) (A, B2) to R7. R7, as Leaf, removes outer IPv6 header and delivers the payload.

A.2.1. Pinging Replication SID

This section illustrates ping of a Replication SID.

Node R1 pings replication SID of node R6 directly by sending the following packet:

1. R1 to R6: (2001:db8::1, 2001:db8:cccc:6:F6::0; NH=ICMPv6) (ICMPv6 Echo Request)

2. Node R6 as a Leaf processes upper layer ICMPv6 Echo Request and responds with ICMPv6 Echo Reply

Node R1 pings Replication SID of R7 via R4 by sending the following packet with SRH:

1. R1 to R4: (2001:db8::1, 2001:db8:cccc:4:C7::0) (2001:db8:cccc:7:F7::0; SL=1; NH=ICMPV6) (ICMPv6 Echo Request)

2. R4 to R7: (2001:db8::1, 2001:db8:cccc:7:F7::0; NH=ICMPv6) (ICMPv6 Echo Request)

3. Node R7 as a Leaf processes upper layer ICMPv6 Echo Request and responds with ICMPv6 Echo Reply

Assume node R4 is a transit Replication node with Replication SID 2001:db8:cccc:4:F4::0 replicating to R7. Node R1 pings Replication SID of R7 via Replication SID of R4 as follows:

1. R1 to R4: (2001:db8::1, 2001:db8:cccc:4:F4::0; NH=ICMPv6) (ICMPv6 Echo Request)

2. R4 replicates to R7 by replacing IPv6 destination address with Replication SID of R7 from its replication state

3. R4 to R7: (2001:db8::1, 2001:db8:cccc:7:F7::0; NH=ICMPv6) (ICMPv6 Echo Request)

4. Node R7 as a Leaf processes upper layer ICMPv6 Echo Request and responds with ICMPv6 Echo Reply

Authors' Addresses

Daniel Voyer (editor)
Bell Canada
Montreal
Canada

Email: daniel.voyer@bell.ca

Clarence Filsfils
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Rishabh Parekh
Cisco Systems, Inc.
San Jose,
United States of America

Email: riparekh@cisco.com

Hooman Bidgoli
Nokia
Ottawa
Canada

Email: hooman.bidgoli@nokia.com

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net