SPRING Working Group
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2025

S. Peng
D. Dhody
Huawei
K. Talaulikar
Cisco Systems
G. Mishra
Verizon Inc.

3 March 2025

Path MTU (PMTU) for Segment Routing (SR) Policy
draft-ietf-spring-pmtu-sr-policy-02

Abstract

This document defines the Path MTU (PMTU) for Segment Routing (SR) Policy (called SR-PMTU). It applies to both Segment Routing over IPv6 (SRv6) and SR-MPLS. This document specifies the framework of SR-PMTU for SR Policy including the link MTU collection, the SR-PMTU computation, the SR-PMTU enforcement, and the handling behaviours on the headend.

Table of Contents

1. Introduction

Segment Routing (SR) [RFC8402] allows a node to steer a packet flow along any given path. The headend is a node where the instructions for source routing (i.e., segments) are encoded in the packet and hence becomes the starting node for a specific segment routing path. Intermediate per-path states are eliminated thanks to source routing.

A Segment Routing Policy (SR Policy) [RFC9256] is an ordered list of segments (i.e., instructions) that represent a source-routed policy. The headend node is said to steer a flow into a SR Policy. The packets steered into an SR Policy have an ordered list of segments associated with that SR Policy written into them. [RFC8660] describes the representation and processing of this ordered list of segments as an MPLS label stack for SR-MPLS, while [RFC8754] and [RFC8986] describe the same for Segment Routing over IPv6 (SRv6) with the use of the Segment Routing Header (SRH).

[RFC8402] introduces the SR Policy construct and provides an overview of how it is leveraged for Segment Routing use-cases. [RFC9256] updates [RFC8402] to specify detailed concepts of SR Policy and steering packets into an SR Policy.

This document extends the SR Policy to also include the Path MTU information to SR Policy and applies to both SRv6 and SR-MPLS. The SRv6-specific handling is specified in Section 6.

1.1. Motivation

The motivation for handling SR-PMTU for the SR paths includes (but is not limited to):

* Being able to avoid fragmentation by being aware of the SR-PMTU associated with the SR paths and policies at the headend.

* Being able to generate ICMP messages at the headend.

* When fragmentation is unavoidable, the ability to do it correctly at the headend.

* Ability to use SR-PMTU as path computation constraint and optimization criteria at the headend or controller/PCE.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

Link MTU: As per [RFC8899], this could more properly be called the IP MTU. It is the size in bytes of the largest IP packet, including the IP header and payload, that can be transmitted over a link. In case of MPLS, it also includes the label stack, and in case of IPv6, it includes IPv6 extension headers (including SRH).

Path MTU (PMTU): See [RFC8899]. In the scope of this document, this is also called SR-PMTU for the SR paths and policies.

SR overhead: The SR-MPLS label stack or SRH. The link MTU takes the SR overhead into consideration.

4. SR-PMTU Definition for SR Policy

The Segment Routing policy architecture is specified in [RFC9256]. An SR Policy is associated with one or more candidate paths. A candidate path is either dynamic, explicit, or composite. The related concepts with the SR-PMTU definition in this document are listed as follows.

An explicit/dynamic candidate path is expressed as a Segment-List or a set of Segment-Lists directly or by computation. If a candidate path is associated with a set of Segment-Lists, each Segment-List is associated with weight for weighted load balancing. The default weight is 1.

A composite candidate path is defined in [RFC9256].

4.1. SR-PMTU of a Segment List

A Segment-List represents a specific source-routed path to send traffic from the headend to the endpoint of the corresponding SR policy [RFC9256]. The SR-PMTU of a segment list is defined as the minimum Link MTU of all the links in a path between a source node and a destination node. Refer to Section 5.2 for specific handling for Node, Adjacency and Binding SID (as well as their combinations).

4.2. SR-PMTU of a Candidate Path

In the case of an explicit/dynamic candidate path, if it is expressed as a single Segment-List, then the SR-PMTU of the candidate path is the same as that of the SR-PMTU of the segment list as described in Section 4.1.

In the case of an explicit/dynamic candidate path, if it is expressed as a set of Segment-Lists (for load-balancing), then the SR-PMTU of the candidate path is defined as the minimum SR-PMTU of all the Segment-Lists in the set.

In the case of a composite candidate path, the SR-PMTU is defined as the minimum SR-PMTU of all the constituent SR Policies.

4.3. SR-PMTU of an SR Policy

According to [RFC9256], an SR Policy is associated with one or more candidate paths. A candidate path is selected when it is valid and it is determined to be the best path of the SR Policy. The selected path is referred to as the "active path" of the SR policy. Then the SR-PMTU for an SR Policy is defined as the SR-PMTU of the selected/active candidate path of this SR policy.

In the case of an explicit/dynamic candidate path, the SR-PMTU definition can be referred to in Section 4.2.

In the case of a composite candidate path, the SR-PMTU is defined as the minimum SR-PMTU of all the constituent SR policies. Since the constituent SR Policies of a composite candidate path can only be explicit/dynamic candidate paths, then the SR-PMTU definition of explicit/dynamic candidate path is as per Section 4.2.

5. The Framework of SR-PMTU for SR Policy

The framework of SR-PMTU for SR Policy includes link MTU collection, SR-PMTU computation, SR-PMTU enforcement, and handling behaviors on the headend.

```
                         +-------------------+
                +--------|Network Controller| SR-PMTU computation
                |        +---------/|\-------+
                |                   |
    SR-PMTU enforcement   Link MTU Collection
                |                   |
             +-\|/-+   +-----------|-----------+   +-----+
   Handling  |Head |---|    Segment Routing    |---|End  |
   behaviors |end  |   |    Network Domain     |   |Point|
             +-----+   +-----------------------+   +-----+
                <---------Link MTU collection---------|
```

Figure 1. The Framework of SR-PMTU for SR Policy

5.1. Link MTU Collection

The SR-PMTU of a segment list is defined as the minimum Link MTU of all the links in a path, see Section 4.1. The Link MTU can be collected in network through various mechanisms such as the ones defined in [I-D.hu-lsr-igp-path-mtu] and [I-D.ietf-idr-bgp-ls-link-mtu] without the knowledge of the services.

5.2. SR-PMTU Computation

The collected link MTU of all the related links are sent to the network controller or the headend where the SR-PMTU is computed. Depending upon the path type, the computation methods are different, which are described in the following subsections.

5.2.1. Loose Path

In a loose path [RFC7855], only Node SIDs are used along the path. Between two adjacent Node-SIDs, generally, there are equal-cost multipaths (ECMP). The SR-PMTU of the loose path is computed by finding the minimum SR-PMTU of all the ECMPs between two adjacent Node SIDs along the loose path.

5.2.2. Strict TE Path

In a strict TE path [RFC7855], only Adj-SIDs are used along the path. Since the link MTU of all the links being indicated by the Adj-SIDs of the strict TE path are known to the network controller, the SR-PMTU of the strict SR-TE path is computed by finding out the minimum link MTU of all the links in the strict SR-TE path between its source node and destination node.

5.2.3. Mixed Path

In a mixed path, both Node SIDs-and Adj-SIDs are used along the path. The PMTU of the mixed TE path is computed by finding the minimum SR-PMTU of all the ECMPs between two adjacent Node SIDs and the link MTU of all the links indicated by the Adj SIDs.

5.2.4. Binding Path

The Binding SID (BSID) [RFC8402] is bound to an SR Policy, instantiation of which may involve a list of SIDs. The SR-PMTU of the binding path is the same as that of an SR Policy as specified in the above section modulo that it also includes the encapsulation overhead associated with it (i.e. the additional label stack pushed in case of SR-MPLS and the outer IPv6 header with its own SRH in case of SRv6). This is done to make sure the headend of the SR path that includes a BSID is able to compute the SR-PMTU correctly by taking the correct SR-PMTU of the binding path into consideration along with other SIDs in the SR path.

5.2.5. TI-LFA

Topology Independent Loop-free Alternate Fast Re-route (TI-LFA) [I-D.ietf-rtgwg-segment-routing-ti-lfa], aims to provide protection for node and adjacency segments within the SR framework. The PMTU of the repair path might be different from the original path's, which could lead to fragmentation while the repair path is in use.

To avoid fragmentation, it is possible for the headend (or controller) to consider the FRR overhead when computing the SR-PMTU of the original path.

5.3. SR-PMTU Enforcement

SR Policy as per [RFC9256] does not include SR-PMTU in the SR Policy encoding structure. As specified in [I-D.ietf-idr-sr-policy-path-mtu], the SR-PMTU is encoded in the SR policy structure as shown in Figure 2. After the SR-PMTU computation, the SR-PMTU is enforced along with the SR Policy to the headend of the corresponding path.

```
SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
Attributes:
   Tunnel Encaps Attribute (23)
      Tunnel Type: SR Policy
          Binding SID
          Preference
          Priority
          Policy Name
          Explicit NULL Label Policy (ENLP)
          Segment List
              Weight
        ----> Path MTU (SR-PMTU)
              Segment
              Segment
              ...
          ...
```

Figure 2. The SR Policy encoding structure with SR-PMTU

When there are multiple paths that can be selected, the one with the highest SR-PMTU will be used in order to avoid fragmentation on the headend.

The PCEP extension to handle PMTU is specified in [I-D.ietf-pce-pcep-pmtu].

5.4. Handling behaviors on the headend

After the SR-PMTU is computed, the headend performs the handling behaviors such as encapsulation and fragmentation, if needed. Note that this behavior is similar to the existing behaviors of MPLS and IPv6 dataplane.

5.4.1. SR-PMTU Constraints and Optimization

Generally, considering the services being carried, the operators set an SR-PMTU constraint aiming for a proper path selection that fulfills packet size requirements hence avoiding fragmentation. Furthermore, the encapsulation on the headend will introduce the overhead on top of the packet to be encapsulated. Generally, the encapsulation overhead has to be estimated according to the possible path hops and sometimes the repair paths. Therefore, the SR-PMTU constraint is set considering both the carried services and the encapsulation overhead.

When SR-PMTU-based path optimization is done, PCE will select the path with the highest SR-PMTU among all the possible paths.

Even if the SR-PMTU is not considered by the PCE at the time of path computation, the computed SR-PMTU is useful at the headend for the reasons already stated in Section 1.1.

Once the SR-PMTU constraint is set on the headend, it is supposed to be the lowest bound of the SR-PMTUs of all the paths being computed locally or enforced by the controller in order to avoid fragmentation.

5.4.2. Fragmentation processing

If the SR-PMTU of all the paths being computed locally or enforced by the controller is smaller than the SR-PMTU constraint set on the headend, the fragmentation will have to be handled. If fragmentation is not possible, the headend could generate the ICMP messages [RFC4443] to notify the traffic source.

Over this selected path, on the headend, the packets are fragmented in order to guarantee the size of the encapsulated packets is smaller than the PMTU of the selected path.

5.5. Link MTU Change

The Link MTU collected as described in Section 5.1 may change over time due to factors such as device configuration updates or topological modifications, such as the addition of a new link with a lower MTU. These changes can impact the SR-PMTU of the data path, and the computed SR-PMTU value may remain outdated until the control plane converges. This behavior is similar to changes in other link metrics.

6. SRv6-Specific Handling

In the case of SRv6, the SRH is included in the calculation of the Link MTU and thus in the SR-PMTU. Note that the PMTU considerations for IPv6 [RFC8201] apply for the SRv6. [RFC8754] also specify the MTU considerations related to encapsulation with an outer IPv6 header with SRH.

7. Security Considerations

[RFC9256] specifies in detail the SR Policy construct (introduced in [RFC8402]) and its security considerations. The additional SR-MTU attribute information can be sensitive in some deployments and could be used to influence SR path setup and selection with adverse effect. The protocol extensions that include SR-PMTU need to take this into consideration. This document does not define any new protocol extensions and thus does not introduce any further security considerations.

8. IANA Considerations

This document does not include any IANA requests.

9. Acknowledgement

Thanks to xx for useful discussions and comments.

10. References

10.1. 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>.

[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>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[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>.

[RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <https://www.rfc-editor.org/info/rfc8201>.

[RFC8899] Fairhurst, G., Jones, T., T端xen, M., R端ngeler, I., and T. V旦lker, "Packetization Layer Path MTU Discovery for Datagram Transports", RFC 8899, DOI 10.17487/RFC8899, September 2020, <https://www.rfc-editor.org/info/rfc8899>.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]
Bashandy, A., Litkowski, S., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, Internet-Draft, draft-ietf-rtgwg-segment-routing-ti-lfa-21, 12 February 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-segment-routing-ti-lfa-21>.

10.2. Informative References

[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>.

[RFC7855] Previdi, S., Ed., Filsfils, C., Ed., Decraene, B., Litkowski, S., Horneffer, M., and R. Shakir, "Source Packet Routing in Networking (SPRING) Problem Statement and Requirements", RFC 7855, DOI 10.17487/RFC7855, May 2016, <https://www.rfc-editor.org/info/rfc7855>.

[I-D.ietf-idr-bgp-ls-link-mtu]
Zhu, Y., Hu, Z., Peng, S., and R. Mwehair, "Signaling Maximum Transmission Unit (MTU) using BGP-LS", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ls-link-mtu-08, 17 September 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ls-link-mtu-08>.

[I-D.hu-lsr-igp-path-mtu]
Hu, Z., Peng, S., and X. Xi, "IGP Extensions for Path MTU", Work in Progress, Internet-Draft, draft-hu-lsr-igp-path-mtu-00, 19 October 2021, <https://datatracker.ietf.org/doc/html/draft-hu-lsr-igp-path-mtu-00>.

[I-D.ietf-idr-sr-policy-path-mtu]
Li, C., Zhu, Y., El Sawaf, A., and Z. Li, "Segment Routing Path MTU in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-sr-policy-path-mtu-10, 24 September 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-sr-policy-path-mtu-10>.

[I-D.ietf-pce-pcep-pmtu]
Peng, S., Li, C., Han, L., Ndifor, L., and S. Sidor, "Support for Path MTU (PMTU) in the Path Computation

Element (PCE) Communication Protocol (PCEP)", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-pmtu-07, 12 February 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-pmtu-07>.

Authors' Addresses

Shuping Peng
Huawei
Huawei Campus, No. 156 Beiqing Rd.
Beijing
100095
China
Email: pengshuping@huawei.com

Dhruv Dhody
Huawei
India
Email: dhruv.ietf@gmail.com

Ketan Talaulikar
Cisco Systems
India
Email: ketant.ietf@gmail.com

Gyan Mishra
Verizon Inc.
United States of America
Email: gyan.s.mishra@verizon.com