BMWG G. Fioccola
Internet-Draft E. Vasilenko
Intended status: Informational P. Volpato
Expires: January 7, 2023 Huawei Technologies
L. Contreras
Telefonica
July 6, 2022

Benchmarking Methodology for MPLS Segment Routing
draft-vfv-bmwg-srmpls-bench-meth-02

Abstract

This document defines a methodology for benchmarking Segment Routing (SR) performance for Segment Routing over MPLS (SR-MPLS). It builds upon [RFC2544], [RFC5695] and [RFC8402].

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], RFC 8174 [RFC8174].

Table of Contents

1. Introduction

Segment Routing (SR), defined in [RFC8402], leverages the source routing paradigm. The headend node steers a packet through an SR Policy [I-D.ietf-spring-segment-routing-policy], instantiated as an ordered list of segments. A segment, referred to by its Segment Identifier (SID), can have a semantic local to an SR node or global within an SR domain. SR supports per-flow explicit routing while

maintaining per-flow state only at the ingress nodes to the SR domain.

However, there is no standard method defined to compare and contrast the foundational SR packet forwarding capabilities of network devices. This document aims to extend the efforts of [RFC1242] and [RFC2544] to SR network.

The SR architecture can be instantiated on two data-plane: SR over MPLS (SR-MPLS) and SR over IPv6 (SRv6). This document is limited to SR-MPLS.

It is expected that future documents may cover the benchmarking of SR-MPLS applications such as Layer 3 VPN (L3VPN) [RFC4364], EVPN [RFC7432], Fast ReRoute [I-D.ietf-rtgwg-segment-routing-ti-lfa], etc.

SR can be directly applied to the Multiprotocol Label Switching (MPLS) architecture with no change to the forwarding plane [RFC8660]. A segment is encoded as an MPLS label. An SR Policy is instantiated as a stack of labels.

SR-MPLS involves 3 types of forwarding plane operations:

- o PUSH consists of the insertion of one or more segments on top of the incoming packet It is the outer label of the SR-MPLS label stack.

- o NEXT consists of the inspection of the next segment. The active segment is completed and the next segment is activated. It is a POP of the top label in SR-MPLS.

- o CONTINUE happens when the active segment is not completed; hence, it remains active. It is a SWAP of the top label in SR-MPLS.

SR list for PUSH operation is typically constructed by SR Policy in ingress node, see [I-D.ietf-spring-segment-routing-policy].

[RFC5695] describes a methodology specific to the benchmarking of MPLS forwarding devices, by considering the most common MPLS packet forwarding scenarios and corresponding performance measurements.

The purpose of this document is to describe a methodology specific to the benchmarking of Segment Routing. The methodology described is a complement for [RFC5695].

2. SR-MPLS Forwarding

In MPLS, a Prefix-SID is allocated in the form of an MPLS label. For SR-MPLS, Segment Routing does not require any change to the MPLS forwarding plane. An SR Policy is instantiated through the MPLS Label Stack: the Segment IDs (SIDs) of a Segment List are inserted as MPLS Labels. The classical forwarding functions available for MPLS networks allow implementing the SR operations.

The operations applied by the SR-MPLS forwarding plane are PUSH, NEXT, and CONTINUE.

The PUSH operation corresponds to the Label Push function, according to the MPLS label pushing rules specified in [RFC3032]. It consists of pushing one or more MPLS labels on top of an incoming packet then sending it out of a particular physical interface or virtual interface towards a particular next hop.

The NEXT operation corresponds to the Label Pop function, which consists of removing the topmost label. The action before and/or after the popping depends on the instruction associated with the active SID on the received packet prior to the popping. It is equivalent to Penultimate Hop Popping (PHP).

The CONTINUE operation corresponds to the Label Swap function, according to the MPLS label-swapping rules in [RFC3031]. It consists of associating an incoming label with an outgoing interface and outgoing label and forwarding the packet to the outgoing interface. It is equivalent to Ultimate Hop Popping (UHP).

The encapsulation of an IP packet into an SR-MPLS packet is performed at the edge of an SR-MPLS domain, reusing the MPLS Forwarding Equivalent Class (FEC) concept. A Forwarding Equivalent Class (FEC) can be associated with an SR Policy ([I-D.ietf-spring-segment-routing-policy]). When pushing labels onto a packet's label stack, the Time-to-Live (TTL) field and the Traffic Class (TC) field of each label stack entry must also be set.

All SR nodes in the SR domain use an IGP signaling extension to advertise their own prefix SIDs. After receiving the advertised prefix SIDs, each SR node calculates the prefix SIDs to the advertisers. The prefix SID advertisement can be an absolute value advertisement or an index value advertisement. In this regard, the mapping of Segments to MPLS Labels (SIDs) is an important process in the SR-MPLS data plane. Each router can advertise its own available label space to be used for Global Segments called Segment Routing Global Block (SRGB) and an identical range of labels (SRGB) should be used in all routers in order to simplify services and operations. In

the SR domain Global Segments can be identified by an index, which has to be re-mapped into a label, or by an absolute value. This is relevant for the nodes that perform the NEXT operation to the segments, because the label for the next segments needs to be crafted accordingly.

[I-D.ietf-spring-segment-routing-policy] specifies the concepts of SR Policy and steering into an SR Policy. The header of a packet steered in an SR Policy is augmented with the ordered list of segments associated with that SR Policy. SR Policy state is instantiated only on the headend node, that steers a flow into an SR Policy. Indeed intermediate and endpoint nodes do not require any state to be maintained. SR Policies can be instantiated on the headend dynamically and on demand basis. SR policy may be installed by PCEP [RFC8664], BGP [I-D.ietf-idr-segment-routing-te-policy], or via manual configuration on the router. PCEP signaling can be the case of a controller based deployment. For all these reasons, SR Policies scale better than traditional TE mechanisms.

3. Test Methodology

3.1. Test Setup

The test setup in general is compliant with section 6 of [RFC2544] but augmented by the methodology specified in section 4 of [RFC5695] using many ports. In fact, it is needed to test the packet forwarding engine that may have different performance based on the number of ports served. The Device Under Test (DUT) may have oversubscribed ports, then traffic for such ports should be proportionally decreased according to the specific DUT oversubscription ratio. All ports served by a particular packet forwarding engine should be loaded in reverse proportion to the claimed oversubscription ratio. Tests SHOULD be done with bidirectional traffic that better reflects the real environment for SR-MPLS nodes. It is OPTIONAL to choose non-equal proportion for upstream and downstream traffic for some specific aggregation nodes.

The recommended topology for SR-MPLS Forwarding Benchmarking should be the same as MPLS and it is described in section 4 of [RFC5695]. Port numbers involved in the tests and their oversubscription ratio MUST be reported. In general, MPLS labels at the bottom of the stack may be used to encode services (L2/L3 VPNs) but it is out of the scope of this document. This document is benchmarking only "source routing". Hence, SIDs represent only prefix and adjacency segments.

Segment Routing may also be implemented as a software network function in an NFV Infrastructure and, in this case, additional considerations should be done. [RFC9004] updates the procedures of

the test to measure the Back-to-Back Frames since their characterization is relevant in software-packet processing. Also, [ETSI-GR-NFV-TST-007] describes test guidelines for NFV capabilities that require interactions between the components implementing NFV functionality.

3.2. Label Distribution Support

As specified in [RFC8402], in the context of an IGP-based distributed control plane, two topological segments are defined: the IGP-Adjacency segment and the IGP-Prefix segment; while, in the context of a BGP-based distributed control plane, two topological segments are defined: the BGP peer segment and the BGP Prefix segment.

It is RECOMMENDED that the DUT and test tool support at least one option for SID stack construction:

- o IS-IS Extensions for Segment Routing [RFC8667]
- o OSPF Extensions for Segment Routing [RFC8665]
- o Segment Routing Prefix Segment Identifier Extensions for BGP [RFC8669]
- o Segment Routing Policy Architecture [I-D.ietf-spring-segment-routing-policy].

It is RECOMMENDED that at least one routing protocol (OSPF or ISIS or BGP) should be used for the construction of the simplest stack of 1 SID. It is RECOMMENDED that SR policy should be used for the construction of a stack with 2 SIDs. It is possible to test longer SID lists if there is an interest.

It is RECOMMENDED that the top SID on the list (outer label) should be an adjacency type to emulate the traffic engineering scenario. In all cases, SID stack configuration SHOULD happen before packet forwarding would be started. Control plane convergence speed is not the subject of the present tests.

The label distribution method and SR policy construction method used MUST be reported according to Section 4.

3.3. Frame Formats and Sizes

The tests for SR-MPLS will use the Frame characteristics similarly to section 4.1.5 of [RFC5695], except the need for a bigger MTU to accommodate many MPLS labels.

It is to be noted that [RFC5695] requires exactly a single entry in the MPLS label stack in an MPLS packet that is not enough to simulate typical SR SID list. MPLS label values used in any test case MUST be outside the reserved label value (0-15) unless stated otherwise. The number of entries in the label stack MUST be reported.

According to section 4.1.4.2 of [RFC5695], the payload is RECOMMENDED to have an IP packet (IPv6 or IPv4 with UDP or TCP) to better represent the real environment.

It is assumed that the test would be for Ethernet media only. Other media is possible (see section 4.1.5.2 of [RFC5695] for the POS example). Recommended frame sizes are presented below. Any other frame sized may be added if suspected of abnormal behavior. For example, some architectures may allocate buffer memory in big fixed chunks that may drop performance if frame sizes are chosen just a few octet more than the fixed chunk size (the second chunk would have a very low memory utilization).

Recommended frame sizes are the following:

- o Ethernet Minimal: 64+n*4
- o DUT Minimal Wire Speed: 128-256 (it depends on the range recommended in the DUT specification)
- o Ethernet Typical: 1518+n*4
- o DUT Maximum Wire Speed: 9000

Note that n*4 octets are added in the previous calculations to accommodate MPLS labels needed for respective tests. The typical frame size values are listed above for the DUT minimal and maximum wire speed, but they can be modified according to the DUT characteristics. Indeed, the minimum wire speed frame size can be considered based on the DUT specification but, in some cases, many tests may be needed in the search for the real minimum wire speed frame size. VLAN tag may additionally increase the frame size. VLAN tag tests are OPTIONAL.

3.4. Protocol Addresses

IANA reserved an IPv6 address block 2001:0200::/48 ([RFC4773]) for use with IPv6 benchmark testing and block 198.18.0.0/15 ([RFC3330]) for IPv4 benchmark testing. The type of infrastructure protocol (IPv6 vs IPv4) that should be used for IGP and BGP in the tests should be chosen according to the test purpose and requirements.

3.5. Trial Duration

The test portion of each trial SHOULD be at least 10 seconds longer than the hold time for the respective protocol configuration to verify that the DUT can maintain a stable control plane when the data-forwarding plane is under stress. IGP protocols typically have a shorter hold time, some BGP default configuration may be up to 180 seconds. It is needed to check the default hold time of the DUT for the respective protocol used.

3.6. Traffic Verification

Traffic verification is following section 4.1.8 of [RFC5695].

4. Reporting Format

There are a few parameters that need to be changed in section 5 of [RFC5695] for SR MPLS tests. New parameters that MUST be reported are:

- o Port numbers involved in the tests and their respective oversubscription ratio.
- o Upstream/downstream traffic proportion (equal bidirectional or some other split).
- o SR-MPLS Forwarding Operations (PUSH/ NEXT/ CONTINUE).
- o Number of Segments considered in the MPLS Label Stack and the type of SIDs used (Global/Local).
- o SR Policy construction method (PCEP, BGP, manual configuration).
- o Type of the payload (IPv6/IPv4, UDP/TCP).

Some parameters MAY be changed:

- o Label Distribution protocol and IGP are the same in the context of SR MPLS. Hence, it is called "label distribution".
- o Port media type may be reported only one time for all tests if only Ethernet media would be tested

5. SR-MPLS Forwarding Benchmarking Tests

In general, tests are compliant with [RFC2544] but the important correction discussed in section 6 of [RFC5695] is applied: ports chosen for every test MUST stress all ports served by one forwarding

engine. It is better to check the DUT specification for the relationship between ports and the forwarding engine to minimize the number of ports involved. But it is possible to understand the worst case by looking at the throughput and latency from the trial tests. If any doubt exists about how full is the offered load for the forwarding engine then it is better to stress all ports of the line card or all ports for the whole router with a centralized forwarding engine. Partial load on forwarding engine would show optimistic results. Controllable traffic distribution between many ports (as specified in section 4 of [RFC5695]) would need separate SID announcements for separate ports. The search for No-Drop Rate (NDR) should be done for every test as explained in section 6 of [RFC5695].

5.1. Throughput

This section contains the description of the tests that are related to the characterization of a DUT's SR-MPLS traffic forwarding throughput.

The list of segments for SR-MPLS is represented as a stack of MPLS labels. There are three distinct operations to be tested: PUSH, NEXT and CONTINUE. These correspond to the three forwarding operations of an MPLS packet: PUSH (or LSP Ingress), POP (or LSP Egress), or SWAP. It is separately discussed only for throughput tests as an example.

5.1.1. Throughput for SR-MPLS PUSH

Objective: To obtain the DUT's Throughput during the PUSH forwarding operation. It is similar to label Push or LSP Ingress forwarding operation, as per section 6.1.1 of [RFC5695].

Procedure: Similar to [RFC5695] with potential extension to test SID list longer than 1 SID (2 are recommended, many are possible).

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

5.1.2. Throughput for SR-MPLS NEXT

Objective: To obtain the DUT's Throughput during the NEXT forwarding operation. It is equivalent to MPLS Label Pop or Penultimate Hop Popping (PHP), as per section 6.1.3 of [RFC5695].

Procedure: Similar to [RFC5695] with potential extension to test SID list longer than 1 SID (2 are recommended, many are possible).

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

5.1.3. Throughput for SR-MPLS CONTINUE

Objective: To obtain the DUT's Throughput during the CONTINUE forwarding operation. It is equivalent to MPLS Label Swap or Ultimate Hop Popping (UHP), as per section 6.1.2 of [RFC5695]. Non-reserved MPLS label values MUST be used.

Procedure: Similar to [RFC5695] with potential extension to test SID list longer than 1 SID (2 are recommended, many are possible).

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

5.2. Latency

Objective: To determine the latency as defined in section 6.2 of [RFC5695] for each of the SR-MPLS forwarding operations (PUSH, NEXT, CONTINUE).

Procedure: Similar to [RFC5695] with potential extension to test SID list longer than 1 SID (2 are recommended, many are possible).

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

5.3. Frame Loss

Objective: To determine the frame-loss rate (as defined in section 6.3 of [RFC5695]) for each of the SR-MPLS forwarding operations of a DUT throughout the entire range of input data rates and frame sizes.

Procedure: Similar to [RFC5695] with potential extension to test SID list longer than 1 SID (2 are recommended, many are possible).

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

5.4. System Recovery

Objective: To characterize the speed at which a DUT recovers from an overload condition for each of the SR-MPLS forwarding operations.

Procedure: Similar to section 6.4 of [RFC5695].

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

5.5. Reset

Objective: To characterize the speed at which a DUT recovers from a device or software reset for each of the SR-MPLS forwarding operations.

Procedure: Similar to section 6.5 of [RFC5695].

Reporting Format: Similar to [RFC5695] with the additional parameters specified in Section 4.

It is OPTIONAL to extend the Reset tests according to [RFC6201] in order to reset only part of the DUT: only line card reset, only process reset (for example ISIS), only one routing engine reset in the configuration with routing engine redundancy, full power interruption, partial power interruption, etc.

6. Security Considerations

Benchmarking methodologies are limited to technology characterization in a laboratory environment, with dedicated address space and constraints. Special capabilities SHOULD NOT exist in the DUT/SUT specifically for benchmarking purposes. Any implications for network security arising from the DUT/SUT SHOULD be identical in the lab and production networks. The benchmarking network topology is an independent test setup and MUST NOT be connected to devices that may forward the test traffic into a production network or misroute traffic to the test management network.

There are no specific security considerations within the scope of this document.

7. IANA Considerations

This document has no IANA actions.

8. Acknowledgements

The authors would like to thank Al Morton for the precious comments and suggestions.

9. References

9.1. Normative References

[RFC1242] Bradner, S., "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, DOI 10.17487/RFC1242, July 1991, <https://www.rfc-editor.org/info/rfc1242>.

[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>.

[RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <https://www.rfc-editor.org/info/rfc2544>.

[RFC3330] IANA, "Special-Use IPv4 Addresses", RFC 3330, DOI 10.17487/RFC3330, September 2002, <https://www.rfc-editor.org/info/rfc3330>.

[RFC4773] Huston, G., "Administration of the IANA Special Purpose IPv6 Address Block", RFC 4773, DOI 10.17487/RFC4773, December 2006, <https://www.rfc-editor.org/info/rfc4773>.

[RFC5695] Akhter, A., Asati, R., and C. Pignataro, "MPLS Forwarding Benchmarking Methodology for IP Flows", RFC 5695, DOI 10.17487/RFC5695, November 2009, <https://www.rfc-editor.org/info/rfc5695>.

[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>.

[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>.

9.2. Informative References

[ETSI-GR-NFV-TST-007]
ETSI, "ETSI GR NFV-TST 007: Network Functions Virtualisation (NFV) Release 3; Testing; Guidelines on Interoperability Testing for MANO", 2020, <https://www.etsi.org/deliver/etsi_gr/NFV-TST/001_099/007/02.06.01_60/gr_nfv-tst007v020601p.pdf>.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-18 (work in progress), June 2022.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]
Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-08 (work in progress), January 2022.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-22 (work in progress), March 2022.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

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

[RFC6201] Asati, R., Pignataro, C., Calabria, F., and C. Olvera, "Device Reset Characterization", RFC 6201, DOI 10.17487/RFC6201, March 2011, <https://www.rfc-editor.org/info/rfc6201>.

[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>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

[RFC9004] Morton, A., "Updates for the Back-to-Back Frame Benchmark in RFC 2544", RFC 9004, DOI 10.17487/RFC9004, May 2021, <https://www.rfc-editor.org/info/rfc9004>.

Authors' Addresses

Giuseppe Fioccola
Huawei Technologies
Riesstrasse, 25
Munich 80992
Germany

Email: giuseppe.fioccola@huawei.com

Eduard Vasilenko
Huawei Technologies
17/4 Krylatskaya str.
Moscow 121614
Russia

Email: vasilenko.eduard@huawei.com

Paolo Volpato
Huawei Technologies
Via Lorenteggio, 240
Milan 20147
Italy

Email: paolo.volpato@huawei.com

Luis Miguel Contreras Murillo
Telefonica
Spain

Email: luismiguel.contrerasmurillo@telefonica.com