Workgroup: Internet Engineering Task Force
Internet-Draft: draft-geib-spring-oam-opt-05
Published: 21 April 2023
Intended Status: Best Current Practice
Expires: 23 October 2023
Authors: R. Geib, Ed.
Deutsche Telekom

An MPLS SR OAM option reducing the number of end-to-end path validations

Abstract

MPLS traceroute implementations validate dataplane connectivity and isolate faults by sending messages along every end-to-end Label Switched Path (LSP) combination between a source and a destination node. This requires a growing number of path validations in networks with a high number of equal cost paths between origin and destination. Segment Routing (SR) introduces MPLS topology awareness combined with Source Routing. By this combination, SR can be used to implement an MPLS traceroute option lowering the total number of LSP validations as compared to commodity MPLS traceroute.

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 23 October 2023.

Copyright Notice

Table of Contents

1. Introduction

Commodity MPLS isn't topology aware and it doesn't support standardized source routing methods. It is reasonable to validate connectivity and locate faults of MPLS LSPs by detecting and testing all existing LSP combinations between a source and a destination node. The source node originates all MPLS echo requests and evaluates all MPLS echo replies. Operational MPLS OAM implementations were present, when SR MPLS entered standardisation. They continue to work reliably in many cases. MPLS domains with a high number of equal cost paths between source and destination nodes push the detection capabilities of commodity MPLS OAM to the limit. So far, modes of MPLS OAM operation adding Segment Routing functionality to deal with limitations of commodity MPLS OAM have not been published within IETF.

This draft assumes readers to be aware of MPLS OAM functionality as specified by RFC 8029 [RFC8029] and RFC 8287 [RFC8287]. The function described in the following works for Shortest Path First Paths or Label stacks based on MPLS Node-SID and MPLS Adj-SIDs (if the latter are distributed by Interior Gateway Protocols).

Networks supporting a high number of equivalent cost paths between source and destination nodes require a high number of completed MPLS path validations. Consider a network with Multiple equal cost paths, as shown in figure 1.

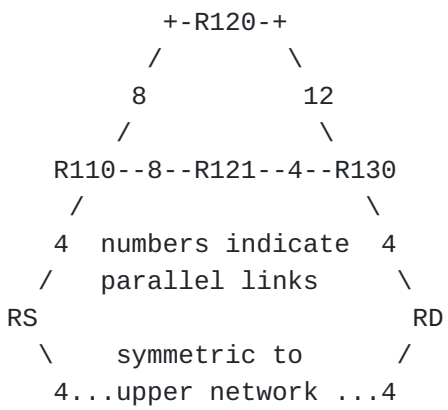

Figure 1

Figure 1: Multiple equal cost path example network.

The total number of MPLS LSP combinations between nodes RS and RD is multiplicative by the number of (equal cost, so to say) links per hop. That results in a maximum of 4096=2*4*(8*12+8*4)*4 path combinations which a commodity MPLS traceroute may try to validate. Assume node RS to start an MPLS traceroute to node RD, containing a Multipath Data Sub-TLV requesting Multipath information for 32 IP-addresses. By Equal Cost Multipath routing (ECMP, [RFC2991]) traffic of likely 16 of these IP-addresses is forwarded via R110 as next hop (the other 16 addresses are assumed to be forwarded along the symmetric and equal cost paths in the lower half of the topology, which are omitted in the figure for brevity). R110 can be expected to respond by an MPLS echo reply indicating prefixes to address each of the 4 equal cost (sub-)paths between RS and R110.

R110 is able to forward traffic addressed by these 16 IP addresses via 16 equal cost paths. There's a fairly high probability that this will not be possible, as some of R110's availble paths to forward traffic to RD will receive traffic of two or even three MPLS echo request destination IP addresses resulting in an MPLS Echo request being sent from RS to R110 and ahead, while other equal cost paths of R110 receive no MPLS traceroute traffic at all. The MPLS Echo Replies returned to RS will indicate that. A commodity solution is, to start an additional MPLS traceroute from RS with another 32 destination IP-addresses. This may help to then enable forwarding of MPLS Echo requests along all of R110's paths to RD via R120 and R121, respectively. With bad luck, R110 will forward only 14 or 15 addresses via R120. R120 forwards MPLS Echo requests along 12 equal cost paths to RD. Then again, there's a fair chance that more destination IP-addresses are required to forward at least one MPLS echo request along all of R120 equal cost paths to RD. Finally, each new MPLS Echo Request containing additional IP destination addresses requires completion of the MPLS Echo-Request / Reply dialogue starting from RS to at least all routers along the path to R120.

In the example, roughly only a fourth of the addresses whose forwarding is validated starting from node RS will be routed via R120. ECMP load balancing "filters away" 75% of the MPLS Echo requests carrying the destination IP-addresses whose forwarding path is to be determined. If however MPLS Echo requests carrying a full set of 32 destination IP-addresses were reaching R120, the probability of being unable to forward at least one MPLS Echo request to each outgoing interface (or path, respectively) at R110 destined to node RD was rather small.

The reason for completing all MPLS Echo Request / Reply dialogues along the path between RS and R120 is figuring out, which destination IP-addresses are routed from R110 to R120 to be available at the latter for local traffic forwarding along paths to RD which can't be addressed otherwise. RFC 8029 section 4.1 'Dealing with Equal-Cost Multipath (ECMP)' concludes, that 'full coverage may not be possible' [RFC8029].

Applying Segment Routing (SR) allows node RS to forward MPLS Echo Request packets with up to, e.g., 32 IP addresses to every node which RS detects on a path to node RD. Doing so reduces the number of local router path options to be checked along the end-to-end paths to no more than the sum of the interfaces belonging to one of the ECMP routes between nodes RS and RD. In the case of the example network above, this sum is 2*(4+8+8+12+4+4)=80 different local router interfaces of routers RS, R110, R120, R121 and R130. That means, that around 2% of the messages and MPLS Label Switched Path checks required with commodity MPLS traceroute implementations are sufficient to validate all local forwarding options for paths from RS to RD (note that the calculation isn't exact, it rather indicates the order of magnitude). The commodity MPLS OAM implementations are neither broken nor not working. SR allows deployment of an additional router local MPLS OAM method to validate high numbers of ECMP routes reliably and fast. The method proposed here reduces the number of MPLS Echo-Request / -Reply dialogues to be stored and completed by the origin node of the path validation and it reduces the number of MPLS Echo-Request / -Reply messages to be processed by intermediate nodes.

The functions specified by this document do not require changes in the MPLS OAM protocol as specified by [RFC8029] and [RFC8287].

1.1. 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].

2. MPLS OAM adding MPLS SR mechanisms

By MPLS Segment Routing (SR), each node of an MPLS SR domain learns this domain's MPLS Node-SID topology [RFC8402]. The SR source routing feature allows to forward packets to each individual node within a SR domain. Combining topology awareness and source routing allows complete validation of all operational intermediate router ECMP path forwarding choices from an RS node to an RD node.

Suppose SR to be deployed in the case of the example network and digits following the letter "R" to indicate the corresponding Node-SIDs. Assume "mixed operation" of commodity MPLS OAM and this draft's proposed option applying SR to direct MPLS echo requests to specific nodes along an end-to-end path. Node RS starts a commodity MPLS Echo request to R110. After having received an MPLS Echo reply from R110 indicating local paths of R110 on which none of the packets with the remaing 16 IP addresses will be forwarded, RS creates an MPLS Echo Request which transports the original 32 IP addresses to R110. To do so, an additional top-Segment is pushed carrying the R110 Node-SID, 110. The message below this additional segment is coded as a standard RFC8287 MPLS Echo request. Two things are special: the TTL of the MPLS header containing the Node SID of RD is always set to 1. Further, a seperate sequence number series needs to be started to distinguish the starting point of this "SR enhanced" MPLS OAM traceroute sequence. Coding space for MPLS OAM Sender's Handle and Sequence Number is sufficient to do that [RFC8029]. If Pen-ultimate Hop Popping (PHP) is active, the R110 Node-SID is implicitly present only on the link to an uplink neighboring node of R110. Still MPLS echo request packets with all 32 IP-destination addresses are forwarded to R110. The chances to address all of the 16 ECMP paths of R110 to RD with the originally configured 32 IP-addresses increase. The same method is repeated for R120. Now the top Segment picked by node RS is the Node-SID of R120, again with a separate Sender's Handle and Sequence Number combination. Note, that the MPLS Echo request destined to R120 doesn't require execution of MPLS OAM functions in R110. Standard SR forwarding applies at R110 and by that the packet is sent to R120. So when the R12x nodes receive their first MPLS echo request, it will contain 32 IP-addresses (which is a significant increase in number of IP adresses as compared to commodity MPLS OAM).

As a result, the MPLS Echo reply tables maintained by RS likely indicate several forwarding masks correlated to the same IP address range (discerned by the intermediate node receiving and responding to each MPLS Echo request with top Segment TTL=1). For every ECMP path at an indermediate node, to which the originating node RS can't foward an MPLS Echo request due to the limited number of available IP-addresses, a suitable SR top segement is added for an additional next MPLS Echo request of node RS. This in the end allows to

circumvent the "IP-address filtering" effect caused by ECMP for standard MPLS OAM packets.

Being able to forward a "complete" set of IP addresses to any interface along an end-to-end path is helpful in locating errors. Enhanced MPLS OAM packet addressing options, as proposed by this draft, also offer more possibilities to test and unambiguosly locate a failed sub-path.

2.1. Operation in an SR MPLS domain applying only IP-header based ECMP

The basic operation is to transport an MPLS Echo request from the sender node sequentially to a next hop identified on any of the paths to a destination node. This is done by applying standard SR methodology, which here consists of pushing one additional Node-SID on top of the Label-stack to be validated by the sender node. The Node-SID is set to the value of the node, whose forwarding plane information is requested by the MPLS Echo request. This is illustrated by figure 2.

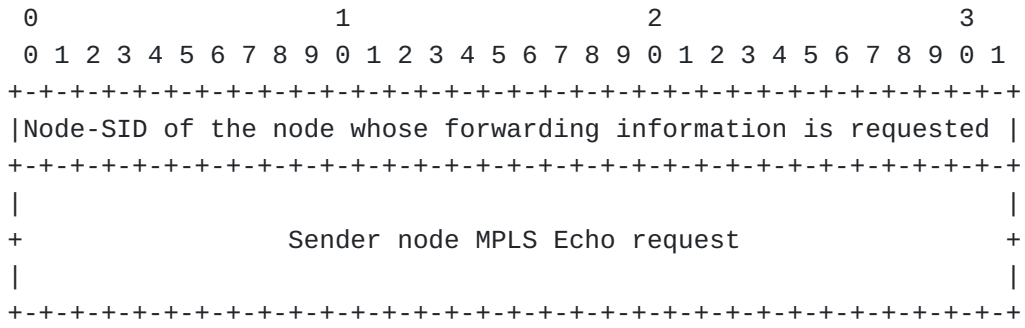

Figure 2

Figure 2: MPLS OAM Label Stack in the case of IP-header only based ECMP.

The added Node-SID is only added to use standard MPLS forwarding. The TTL of this added Node-SID set to the default value for traffic injected by the sending router. The MPLS-TC may be set to a value ensuring reliable transport up to the node, whose forwarding information is requested by the sender node (be aware of MPLS-TC treatment of the node popping this added Node-SID in that case).

The TTL of the top Label of the sender node MPLS Echo request which is contained below the added Node-SID initially is set to TTL=1. Other TTL values can be picked if LSPs from the intermediate node onwards to the destination node of that FEC are desired to be traced or pinged by MPLS OAM messages.

Two modes of operation exist: either applying legacy MPLS OAM and adding the described functionality as required or only applying the

option specified here. Note that the exact path from the sender node to the intermediate node identified by the pushed Node-SID is only known to the node originating and maintaining the MPLS traceroute information, if only one path exists between that sender node and an intermediate node.

If the method is added to commodity MPLS OAM functions, the originatior IP-address of an MPLS Echo-reply indicating a lack of IP-addresses to forward traffic along all ECMP egress interfaces at that intermediate node can be used to derive the Node-SID to be pushed by the MPLS Echo request sender node.

2.2. Operation in an SR MPLS domain additionally using incoming interface information for ECMP

This option can only be applied, if the Segment Routing domain's Adj-SID topology is known to the node originating MPLS Echo Request messages. Configuring the the Interior Gateway Protocol to distribute Adj-SIDs conveniently enables that. If ECMP is additionally using the incoming interface of a packet for path selection, an Adj-SID is added between the Node-SID and the MPLS Echo request. As the idea is to determine the incoming interface of the node, whose ECMP path choices are requested by MPLS OAM, the additionaly pushed Node-SID here is that of the node preceding the intermediate node, whose forwarding information is requested. The Adj-SID is chosen to correspond to a specific incoming interface of the intermediate node whose forwarding information is requested. As the aim of that test is to ensure that every incoming to outgoing interface path choice of the intermediate node can be addressed, the topology information required to identify the upstream Adj-SID corresponding to an incoming interface of the intermediate node is assumed to be present at and maintained by the node originating the MPLS data plane failure test. This additional MPLS to IP topology information excerpt results from prior MPLS path validations of the same basic set of MPLS path validations between the source node and the destination node (this is to express, that no extra measurement effort is caused, as correlation of available information is sufficient). The resulting label stack is illustrated by figure 3.

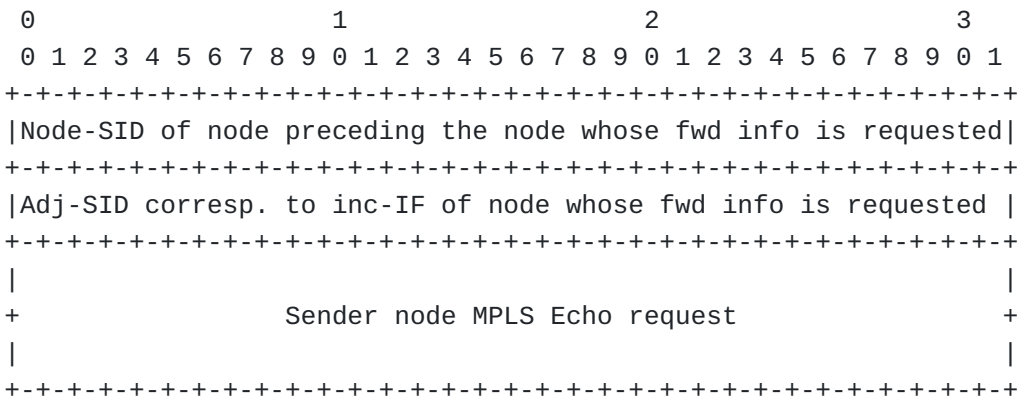

Figure 3

Figure 3: MPLS OAM Label Stack applying SR features if ECMP is additionally based on incoming interfaces.

In the network example of figure 1, node RS picks the Node-SID of R110 and an Adj-SID of R110 corresponding to a particular incoming interface of R120, if the latter's ECMP path also depends on the incoming interface, by which the MPLS Echo request was received.

Here, the full set of original IP-addresses can be forwarded individually per incoming interface of the router whose MPLS forwarding information is requested. In the example above, it is node R120 (not node R110.) Monitoring incoming interface based ECMP results in a higher number of MPLS OAM validations, no matter whether commodity MPLS OAM is applied or the option specified here. The overall sum of tests now is determined by the sum of per node incoming * outgoing paths (or interfaces, respectively). If the method specified here is applied in the case of the example network, 2*(4*8 + 4*8 + 8*12 + 8*4 + 12*4 + 4*4) = 512 MPLS Echo-Request / Response validations are required. Note that this is still a smaller number as compared to the original 4096 path validations resulting in the case of comodity MPLS OAM based on IP-address information only deployed by a domain applying ECMP. Note that the number of required MPLS OAM path validations is increasing significantly, if ECMP forwarding is in addition based on incoming interfaces and the product of a nodes incoming * outgoing interfaces is high.

2.3. Backwards compatibility

This document proposes to add standard Segment Routing functionality to a node originating and controlling MPLS traceroute operation to a destination node. Any changes of the standard MPLS operation only apply there. All other nodes including the destination node don't have to be updated. This allows for a smooth upgrade of an SR domain, starting maybe just with a single node supporting the feature specified here to test and gain experience with MPLS OAM

enhanced by SR functionality and compare operation to commodity MPLS OAM.

3. IANA Considerations

This memo includes no request to IANA.

4. Security Considerations

This document does not introduce new functionality. The approach proposed tries to optimise existing and working implementations. To do so, it combines Segment Routing functions with those of MPLS OAM. These are intra domain functions, and no new attack paths are offered, as changes apply to topology-awareness and addressing options along a path which is addressed by MPLS OAM anyway. No new protocol functions are introduced. The related security sections of both original standards apply, see [RFC8029] and [RFC8402].

5. References

5.1. 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>.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, DOI 10.17487/RFC2991, November 2000, <https://www.rfc-editor.org/info/rfc2991>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Kumar Nainar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8287] Kumar Nainar, N., Pignataro, C., Swallow, G., Akiya, N., Kini, S., and M. Chen, "Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) IGP-Prefix and IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017, <https://www.rfc-editor.org/info/rfc8287>.

[RFC8402] Filsfils, C., Previdi, S., 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>.

Author's Address

Ruediger Geib (editor)
Deutsche Telekom
Ida-Rhodes-Str. 2
64295 Darmstadt
Germany

Phone: +49 6151 5812747
Email: Ruediger.Geib@telekom.de