6MAN
Internet Draft
Intended status: Informational
Expires: January 2, 2018

Z. Ali
C. Filsfils
N. Kumar
F. Iqbal
Cisco Systems, Inc.
R. Raszuk
Bloomberg LP
B. Peirens
Proximus
G. Naik
Drexel University

July 2, 2017

Operations, Administration, and Maintenance (OAM) in Segment Routing Networks with IPv6 Dataplane (SRv6)

draft-ali-6man-srv6-oam-00.txt

Abstract

This document outlines various use-cases for Operations, Administration, and Maintenance (OAM) in Segment Routing with the IPv6 data plane (SRv6) network. It also describes how the existing OAM mechanisms can be used to address SRv6 OAM requirements.

Table of Contents

1. Introduction..2
   1.1. Terminology and Reference Topology........................3
2. Use-cases...4
   2.1. Connectivity Verification.................................4
   2.2. Monitoring A Specific Flow................................5
   2.3. Monitoring all ECMP/ UCMP Paths...........................5
   2.4. Proof of Transit..5
   2.5. Detecting Path Divergence.................................6
   2.6. Fault Isolation...6
   2.7. Centralized OAM...6
3. OAM Mechanisms..6
   3.1. ICMPv6 Applicability......................................6
      3.1.1. Ping...7
      3.1.2. Error Reporting......................................8
      3.1.3. Traceroute...8
   3.2. In-situ OAM..10
   3.3. Seamless BFD Applicability...............................10
   3.4. Controller based OAM.....................................11
4. Security Considerations..12
5. IANA Considerations..12
6. References...12
   6.1. Normative References.....................................12
   6.2. Informative References...................................12
7. Acknowledgments..12

1. Introduction

This document outlines various SRv6 OAM use-cases. It also describes how the existing OAM mechanisms can be used to address SRv6 OAM requirements.

Additional OAM use-cases and mechanisms will be added in a future revision of the document.

1.1. Terminology and Reference Topology

This document uses the terminology defined in [I-D.filsfils-spring-srv6-network-programming]. The readers are expected to be familiar with the same.

Throughout the document, the following simple topology is used for illustration.

```
               |--------------------| N100 |------------------|
               |                                              |
            ====== link1====== link3------ link5====== link9------
            ||N1||======||N2||======| N3 |======||N4||======| N5 |
            ||  ||------||  ||------|    |------||  ||------|    |
            ====== link2====== link4------ link6======link10------
                          |                       |
                          |       ------          |
                          +-------| N6 |---------+
                            link7 |    | link8
                                  ------
```

Reference Topology

In the reference topology:
All nodes are internal nodes within a single SRv6 domain of trust
Nodes N1, N2, and N4 are SRv6 capable nodes.

Nodes N3, N5 and N6 are classic IPv6 nodes.

Node 100 is an SRv6 capable node that acts as controller.

Node Nk has a classic IPv6 loopback address Bk::/128

Node Nk has Ak::/48 for its local SID space from which Local SIDs are explicitly allocated.

The IPv6 address of the nth Link between node X and Y at the X side is represented as 99:X:Y::Xn. e.g.,the IPv6 address of link6 (the 2nd link) between N3 and N4 at N3 in Figure 1 is 99:3:4:32. Similarly, the IPv6 address of link5 (the 1st link between N3 and N4) at node 3 is 99:3:4:31.

Ak::0 is explicitly allocated as the END function at Node k.

Ak::Cij is explicitly allocated as the END.X function at node k towards neighbor node i via jth Link between node i and node j.

e.g., A2::C31 represents END.X at N2 towards N3 via link3 (the 1st link between N2 and N3). Similarly, A4::C52 represents the END.X at N4 towards N5 via link10.

SRH is the abbreviation for the Segment Routing Header.

SL is the abbreviation for the Segment Left.

SID is the abbreviation for the Segment ID.

<S1, S2, S3> represents a SID list where S1 is the first SID and S3 is the last SID. (S3, S2, S1; SL) represents the same SID list but encoded in the SRH format where the rightmost SID (S1) in the SRH is the first SID and the leftmost SID (S3) in the SRH is the last SID.

ECMP is the abbreviation for the Equal Cost Multi-Path.

UCMP is the abbreviation for the Unequal Cost Multi-Path.

2. Use-cases

This section outlines some for the basic OAM use-cases in an SRv6 network. Additional use-cases will be added in a future revision of the document.

2.1. Connectivity Verification

One of the basic OAM use-cases for any network is the capability to perform path monitoring between different end points over any possible shortest path without any path preference. Such essential path monitoring helps to monitor the path availability and the liveliness of the remote end point.

The shortest path monitoring can be done continuously or can be triggered on demand basis using an external event like a script or a CLI trigger. It may be required to perform the connectivity verification in the order of milliseconds, or at a slower pace.

In the reference topology in Figure 1, N1 can send OAM probe packet destined to loopback address of N5 (B5::) to monitor the path liveliness between N1 and N5. N1 optionally may include any relevant segment list in SRH. N1 is not concerned about which route is taken by the probe between N1 and N5 as long as N1 receives the response back from N5. All transit nodes treat the probe packet as like other data packet and forward it based on the Destination Address (DA). N5 looks into the payload of probe packet and respond back to the source address of the probe packet (N1).

2.2. Monitoring A Specific Flow

The network OAM needs to have the ability to monitor a particular path from the available ECMP paths. For example, in the reference topology in figure 1, there are many ECMP paths between N1 and N5. However, the service provider may like to monitor a flow that follows [N1]<link1>[N2]<link7>[N6]<link8>[N4]<link9>[N5].

The flow monitoring can be done continuously or can be triggered on demand basis. It may be required to perform the connectivity verification in the order of milliseconds, or at a slower pace.

2.3. Monitoring all ECMP/ UCMP Paths

In any network, it is common to see multiple ECMP paths between end points that are used for load balancing or redundancy. While monitoring, the shortest path helps to monitor the path and liveliness of remote node, it may not be sufficient to detect any failure in one of the ECMP paths. In our reference topology in figure 1, N6 has 2 ECMP paths to reach N5 as below:

N6--<link8>--N4--<link9>--N5

N6--<link8>--N4--<link10>--N5

If the probe packet from N6 to N5 uses link10, it may not detect any failure on link9. It is critical and beneficial to discover and monitor all ECMP/ UCMP paths. Monitoring of all ECMP/ UCMP paths can be done by probing the candidate paths from end-to-end or by each node by monitoring its data plane.

2.4. Proof of Transit

Various scenarios require the packet to be steered over a particular links or nodes. For example:

- Voice traffic in a SLA constrained network needs to traverse a low latency path between endpoints which may not be the shortest path, i.e. the voice traffic needs to be traffic engineered and steered over the specified segment list that satisfies the SLA constraint.

- In a service chaining environment, the traffic may need to traverse over an ordered list of service functions.

In these scenarios, the SRH contains the list of SID functions that the packet should execute before reaching the destination. It is

possible, due to an error, that the packet may reach the destination without visiting all the segments in the segment list. It is, therefore, important to have the ability to verify that all the function SIDs have been executed correctly before the packet is delivered to the destination. It is also important to ensure that the order of execution of the SID function has been consistent with the SRH contents.

2.5. Detecting Path Divergence

Path divergence occurs when network traffic diverges from the expected path that packet was supposed to take. Path divergence may result in congestion, delay, or breakage of strict SLAs promised to customers. It is, therefore, important to exercise mechanisms that can detect path divergence in the SRv6 network.

2.6. Fault Isolation

In the cases where a monitoring technique discovers an issue, it is required to have the ability to pinpoint the failure location. The fault isolation mechanisms are required to help service providers troubleshoot failure in an SRv6 network.

2.7. Centralized OAM

In the recent past, network operators are interested in performing network operations, administration, and maintenance configuration in a centralized manner. In this use-case, one of the requirements is to implement centralized OAM functionality without any control plane intervention at the monitored nodes.

Additional OAM use-cases will be included in a future revision of the document.

3. OAM Mechanisms

This section describes how existing OAM mechanisms can be used in an SRv6 network. Additional OAM mechanisms will be added in a future revision of the document.

3.1. ICMPv6 Applicability

[RFC4443] describes Internet Control Message Protocol for IPv6 (ICMPv6) that is used by IPv6 devices for network diagnostic and error reporting purposes. As Segment Routing with IPv6 data plane (SRv6) simply adds a new type of Routing Extension Header, existing ICMPv6 mechanisms can be used in an SRv6 network. This section

describes the applicability of ICMPv6 in the SRv6 network and how the existing ICMPv6 mechanisms can be used for basic OAM functionality to address many use-cases outlined in Section 2.

Throughout this document, unless otherwise specified, the acronym ICMPv6 refers to multi-part ICMPv6 messages [RFC4884]. The document does not propose any changes to the standard ICMPv6 [RFC4443], [RFC4884] or standard ICMPv4 [RFC792].

3.1.1. Ping

There is no change required for ping operation at the classic IPv6. Similarly, the existing ping mechanism works along the IGP shortest paths at an SRv6 capable node. However, if an SRv6 capable ingress node wants to ping an IPv6 prefix via an arbitrary segment list <S1, S2, S3>, it needs to initiate ICMPv6 ping with an SR header containing the SID list <S1, S2, S3>. The originator can appropriately set the flow-label field in the IPv6 header of the echo request to influence Equal-Cost Multi-Path (ECMP).

Figure 2 contains sample output for a ping request initiated at node N1 to the loopback address of node N5 via a segment list <A2::C31, A4::C52>.

```
> ping B5:: via segment-list A2::C31, A4::C52

Sending 5, 100-byte ICMP Echos to B5::, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 0.625
/0.749/0.931 ms
```

A sample ping output at an SRv6 capable node

All transit nodes process the echo request message like any other data packet carrying SR header and hence do not require any change. Similarly, the egress node (IPv6 classic or SRv6 capable) does not require any change to process the ICMPv6 echo request. For example, in the ping example of Figure 2:

- Node N2, which is an SRv6 capable node, performs the standard SRH processing. Specifically, it executes the END.X function (A2::C31) on the echo request packet.
- Node N3, which is a classic IPv6 node, performs the standard IPv6 processing. Specifically, it forwards the echo request based on DA A4::C52 in the IPv6 header.
- Node N4, which is an SRv6 capable node, performs the standard SRH processing. Specifically, it observes the END.X function (A4::C52)with PSP (Penultimate Segment POP) on the echo request

packet and removes the SRH and forwards the packet across link10 to N5.
- The echo request packet at N5 arrives as an IPv6 packet without a SRH. If the SRH arrives at classic N5, with SL=0, it should ignore the routing header and process normally. Node N5, which is a classic IPv6 node, performs the standard IPv6/ ICMPv6 processing on the echo request.

3.1.2. Error Reporting

Any IPv6 node can use ICMPv6 control messages to report packet processing errors to the host that originated the datagram packet. To name a few such scenarios:

- If the router receives an undeliverable IP datagram, or
- If the router receives a packet with a Hop Limit of zero, or
- If the router receives a packet such that if the router decrements the packet's Hop Limit it becomes zero, or
- If the router receives a packet with problem with a field in the IPv6 header or the extension headers such that it cannot complete processing the packet, or
- If the router cannot forward a packet because the packet is larger than the MTU of the outgoing link.

In the scenarios listed above, the ICMPv6 response also contains the IP header, IP extension headers and leading payload octets of the "original datagram" to which the ICMPv6 message is a response. Specifically, the Destination Unreachable Message, Time Exceeded Message, Packet Too Big Message and Parameter Problem Message ICMPV6 messages can contain as much of the invoking packet as possible without the ICMPv6 packet exceeding the minimum IPv6 MTU [RFC4443], [RFC4884]. In an SRv6 network, the copy of the invoking packet contains the SR header. The packet originator can use this information for diagnostic purposes. For example, traceroute can use this information as detailed in the following.

3.1.3. Traceroute

There is no change required for traceroute operation at the classic IPv6. Similarly, the existing ping mechanism works along the IGP shortest paths at an SRv6 capable node. However, if an SRv6 capable ingress node wants to traceroute to IPv6 prefix via an arbitrary segment list <S1, S2, S3>, it needs to initiate traceroute probe with an SR header containing the SID list <S1, S2, S3>. The originator can appropriately set the flow-label field in the IPv6 header of the traceroute probe to influence Equal-Cost Multi-Path (ECMP).

Figure 3 contains sample output for a traceroute request initiated at node N1 to the loopback address of node N5 via a segment list < A2::C31, A4::C52>.

```
> traceroute B5:: via segment-list A2::C31, A4::C52

Tracing the route to B5::

 1  99:1:2::21 0.512 msec 0.425 msec 0.374 msec
    SRH: (B5::, A4::C52, A2::C31, SL=2)

 2  99:2:3::31 0.721 msec 0.810 msec 0.795 msec
    SRH: (B5::, A4::C52, A2::C31, SL=1)

 3  99:3:4::41 0.921 msec 0.816 msec 0.759 msec
    SRH: (B5::, A4::C52, A2::C31, SL=1)

 5  99:4:5::52 0.879 msec 0.916 msec 1.024 msec
```

A sample traceroute output at an SRv6 capable node

Please note that information for hop2 is returned by N3, which is a classic IPv6 node. Nonetheless, the ingress node is able to display SR header contents as the packet travels through the IPv6 classic node. This is because the "Time Exceeded Message" ICMPv6 message can contain as much of the invoking packet as possible without the ICMPv6 packet exceeding the minimum IPv6 MTU [RFC4443]. The SR header is also included in these ICMPv6 messages initiated by the classic IPv6 transit nodes that are not running SRv6 software. Specifically, a node generating ICMPv6 message containing a copy of the invoking packet does not need to understand the extension header(s) in the invoking packet.

The segment list information returned for hop1 is returned by N2, which is an SRv6 capable node. Just like for hop2, the ingress node is able to display SR header contents for hop1.

There is no difference in processing of the traceroute probe at an IPv6 classic node and an SRv6 capable node. Similarly, both IPv6 classic and SRv6 capable nodes use the address of the interface on which probe was received as the source address in the ICMPv6 response. ICMP extensions defined in [RFC5837] can be used to also display information about the IP interface through which the datagram would have been forwarded had it been forwardable, and the IP next hop to which the datagram would have been forwarded, the IP interface upon which a datagram arrived, the sub-IP component of an IP interface upon which a datagram arrived.

The information about the IP address of the incoming interface on which the traceroute probe was received by the reporting node is very useful. This information can also be used to verify if SID functions A2::C31 and A4::C52 are executed correctly by N2 and N4, respectively. Specifically, the information displayed for hop2 contains the incoming interface address 99:2:3::31 at N3. This matches with the expected interface bound to END.X function A2::C31 (link3). Similarly, the information displayed for hop5 contains the incoming interface address 99:4:5::52 at N5. This matches with the expected interface bound to the END.X function A4::C52 (link10).

3.2. In-situ OAM

[I-D.brockners-inband-oam-requirements] describes motivation and requirements for In-situ OAM (iOAM). iOAM records operational and telemetry information in the data packet while the packet traverses the network of telemetry domain. iOAM complements out-of-band probe based OAM mechanisms such ICMP ping and traceroute by directly encoding tracing and the other kind of telemetry information to the regular data traffic.

[I-D.brockners-inband-oam-transport] describes transport mechanisms for iOAM data including IPv6 and Segment Routing traffic. furthermore, [I-D.brockners-inband-oam-data] defines information encoding for iOAM data.

One of the applications of iOAM is to provide the Proof of Transit (POT). Among other features of iOAM, SRv6 networks can use the POT feature of iOAM to verify that all the function SIDs in SRH have been executed before the packet is delivered to the destination. It can also ensure that the order of execution of the SID function has been consistent with the SRH contents.

More details on various applications of iOAM in SRv6 networks will be included in future versions of this document.

3.3. Seamless BFD Applicability

[RFC7880] defines Seamless BFD (S-BFD) architecture that simplifies BFD mechanism and enables it to perform path monitoring in a controlled and scalable manner. [RFC7881] describes the procedure to perform continuity check using S-BFD in different environments including IPv6 networks. Section 5.1 of [RFC7881] explains the SBFDInitiator specification and procedure to initiate S-BFD control packet in IP and MPLS network. The specification described for IP-routed S-BFD control packet is also directly applicable to the SRv6 network.

S-BFD has a fast bootstrapping capability. Furthermore, in S-BFD, only the ingress is required to keep BFD states; the egress and transit node does not have any knowledge of the BFD session. These attributes of S-BFD makes it an excellent candidate for rapid failure detection in the SRv6 network. More details on various S-BFD usage on the SRv6 network will be included in a future version.

3.4. Controller based OAM

In the recent past, network operators are interested in performing network operations, administration, and maintenance configuration in a centralized manner. Various data models like YANG are available to collect data from the network and manage it from a centralized entity.

SR technology enables a centralized OAM entity to perform path monitoring from centralized OAM entity without control plane intervention on monitored nodes. [I.D-draft-ietf-spring-oam-usecase] describes such a centralized OAM mechanism. Specifically, the draft describes a procedure that can be used to perform path continuity check between any nodes within an SR domain from a centralized monitoring system, with minimal or no control plane intervene on the nodes. However, the draft focuses on SR networks with MPLS data plane. The same concept applies to the SRv6 networks. This document describes how the concept can be used to perform path monitoring in an SRv6 network.

In the above reference topology, N100 is the centralized monitoring system implementing an END function A100::. In order to verify a segment list <A2::C31, A4::C52>, N100 generates a probe packet with SRH set to (A100::, A4::C52, A2::C31, SL=2). The controller routes the probe packet towards the first segment, which is A2::C31. N2 performs the standard SRH processing and forward it over link3 with the DA of IPv6 packet set to A4::C52. N4 also performs the normal SRH processing and forward it over link10 with the DA of IPv6 packet set to A100::. This makes the probe loops back to the centralized monitoring system. Please note that there is no control plane intervention at the monitored nodes. The entire data plane is exercised at the monitored nodes.

In our reference topology in Figure 1, N100 uses an IGP protocol like OSPF or ISIS to get the topology view within the IGP domain. N100 can also use BGP-LS to get the complete view of an inter-domain topology. In other words, the controller leverages the visibility of the topology to monitor the paths between the various endpoints without control plane intervention required at the monitored nodes.

4. Security Considerations

This document does not define any new protocol extensions and relies on existing procedures defined for ICMP. This document does not impose any additional security challenges to be considered beyond security considerations described in [RFC4884], [RFC4443], [RFC792] and RFCs that updates these RFCs.

5. IANA Considerations

This document does not define any new protocol or any extension to an existing protocol.

6. References

6.1. Normative References

[RFC4884] Extended ICMP to Support Multi-Part Messages. R. Bonica, D. Gan, D. Tappan, C. Pignataro. April 2007.

[RFC4443] Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification. A. Conta, S. Deering, M. Gupta, Ed. March 2006.

[RFC792] Internet Control Message Protocol. J. Postel. September 1981.

[RFC5837] Extending ICMP for Interface and Next-Hop Identification. A. Atlas, Ed., R. Bonica, Ed., C. Pignataro, Ed., N. Shen, JR. Rivers. April 2010.

[RFC7880] Seamless Bidirectional Forwarding Detection (S-BFD). C.Pignataro, D.Ward, N.Akiya, M.Bhatia, S.Pallagatti. July 2016.

[RFC7881] Seamless Bidirectional Forwarding Detection (S-BFD) for IPv4, IPv6, and MPLS. C.Pignataro, D.Ward, N.Akiya. July 2016.

[RFC7880] Seamless Bidirectional Forwarding Detection (S-BFD). C.Pignataro, D.Ward, N.Akiya, M.Bhatia, S.Pallagatti. July 2016.

[RFC7881] Seamless Bidirectional Forwarding Detection (S-BFD) for IPv4, IPv6, and MPLS. C.Pignataro, D.Ward, N.Akiya. July 2016.

[I.D-draft-ietf-spring-oam-usecase] A Scalable and Topology-Aware MPLS Dataplane Monitoring System. R. Geib, C. Filsfils, C. Pignataro, N. Kumar.

[I-D.brockners-inband-oam-data] Data Formats for In-situ OAM. F. Brockners, work in progress.

[I-D.brockners-inband-oam-transport] Encapsulations for In-situ OAM Data, F.Brockners, work in progress.

6.2. Informative References

[I-D.filsfils-spring-srv6-network-programming] SRv6 Network Programming, draft-filsfils-spring-srv6-network-programming, C. Fisfils, et al.

[I-D.brockners-inband-oam-requirements] Requirements for In-situ OAM, F.Brockners, work in progress.

7. Acknowledgments

To be added.

Authors' Addresses

Clarence Filsfils
<Cisco Systems, Inc.>
Email: cfilsfil@cisco.com

Zafar Ali
Cisco Systems, Inc.
Email: zali@cisco.com

Nagendra Kumar
Cisco Systems, Inc.
Email: naikumar@cisco.com

Faisal Iqbal
Cisco Systems, Inc.
Email: faiqbal@cisco.com

Robert Raszuk
Bloomberg LP
731 Lexington Ave
New York City, NY10022, USA
Email: robert@raszuk.net

Bart Peirens
Proximus
Netherlands
Email: bart.peirens@proximus.com

Gaurav Naik
Drexel University
United States of America
Email: gn@drexel.edu