Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: December 09, 2013

N. Akiya
C. Pignataro
D. Ward
Cisco Systems

June 07, 2013

Seamless Bidirectional Forwarding Detection (BFD) with MPLS Label Verification Extension
draft-akiya-bfd-seamless-base-00

Abstract

This specification defines a generic simplified mechanism to use Bidirectional Forwarding Detection (BFD) with large portions of negotiation aspects eliminated, and to allow full and partial reachability validations. For MPLS based BFD, extensions to the generic mechanism are defined for BFD to perform a level of label verifications.

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].

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 http://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 December 09, 2013.

Table of Contents

1. Introduction

Bidirectional Forwarding Detection (BFD), [RFC5880] and related documents, has efficiently generalized the failure detection mechanism for multiple protocols and applications. There are some improvements which can be made to better fit existing technologies. There are also possibility of evolving BFD to better fit new technologies. This document focuses on several aspects of BFD in order to further improve efficiency, to expand failure detection coverage and to allow BFD usage for wider scenarios.

- o There are scenarios where only one side of the BFD, not both, is interested in performing reachability validations. One example is when static route uses BFD to validate nexthop IP address. Another example is when uni-directional tunnel uses BFD to validate reachability to egress node. With these scenarios, corresponding BFD sessions need to be provisioned or instantiated on target nodes but adds very minimal value to those nodes, if any.

- o It is expected that some MPLS technologies will require traffic engineered LSPs to get created dynamically, driven by external applications (ex: SDN). If BFD was to perform reachability validations on LSP prior to requested network node communicating back to the application of LSP readiness, then it would be desirable for BFD to come up fast in order to allow requesting application to proceed. [RFC5884] can take some time as BFD sessions need to get created on both ends, egress via LSP ping, and then session negotiations take place.

- o Existing BFD mechanics provides end-to-end reachability validations well. It does not, however, allow BFD to perform partial reachability validations: ingress to transit, transit to transit, transit to egress.

- o [RFC5884] defines a mechanism to run BFD on exiting MPLS technologies. This mechanism is very useful to perform end-to-end LSP liveliness verification check. However, this mechanism lacks the ability to validate traversal of the intended LSP path. Specifically when one or more nodes along the LSP incorrectly label switch the BFD packet, but it still manages to reach the intended LSP egress node. Likelihood of this issue being seen depends on deployed MPLS technologies. With MPLS technologies which make use of downstream label allocation scheme (ex: RSVP, LDP), incoming label itself provides a level of check as a node will drop any packet containing non-self-advertised label as the top label or will get delivered to unintended egress node. For those MPLS technologies, issue will be less likely to be seen.

With MPLS technologies such as Segment Routing (SR), incoming label can often be a label allocated and advertised by a node that is multiple downstream hops away. For such MPLS technologies, issue will be more likely to be seen. [RFC4379] can detect such broken LSPs, but it is often difficult to run this technology at the rate which BFD is capable of.

- o A node may desire to run multiple BFD sessions to a network target. One such scenario is if multiple applications on the system required to run BFD to a same target but with different failure detection time requirements. Another scenario is to run multiple instances of BFD, hosted on different parts of the system (ex: different CPUs), to a same network target, in order to increase BFD failure reliability by reducing the chance of unrelated local fault causing BFD to declare failure.

This specification provides solutions to above aspects by defining a generic simplified mechanism to use Bidirectional Forwarding Detection (BFD) with large portions of negotiation aspects eliminated, and to allow full and partial reachability validations. For MPLS based BFD, extensions to the generic mechanism are defined for BFD to perform a level of label verifications.

The reader is expected to be familiar with the BFD, IP, MPLS and SR terminologies and protocol constructs.

2. Seamless BFD Overview

Seamless BFD creates packet reflection points in the network. A network node is able to send BFD control packets to these reflection points, and expect response BFD control packets. These reflection points are called BFD target identifiers. Pseudo BFD session instances, referred to as reflector BFD sessions, created on BFD target identifiers are responsible for processing "ping" BFD control packets and generating "pong" BFD control packets.

3. BFD Target Identifier Types

Number of network identifiers types (ex: IP address, segment ID) can make use of this mechanism. To differentiate between different network identifier types, a value is assigned to each type.

BFD Target Identifier types:

Value	BFD Target Identifier Type
0	Reserved
1	IP (IPv4 Address and Router ID)

2 Segment Routing Node Segment ID

Note that IP based BFD from [RFC5885] is supported by this specification, but IP-less based BFD is outside the scope of this document.

Further identifier types to be defined as needed basis.

4. Reserved BFD Discriminators

All local network identifiers which are to participate in this mechanism are to have specific BFD discriminators assigned. Assigned BFD discriminators are attached to corresponding identifiers until they are explicitly un-provisioned. BFD discriminators used for this mechanism are considered reserved, and MUST NOT be reused for other BFD sessions.

Some examples of network identifier to BFD discriminator mappings:

- BFD Target Identifier Type 1: IPv4 address 1.1.1.1 maps to BFD discriminator 0x01010101.
- BFD Target Identifier Type 2: Node segment ID 0x03E800FF maps to BFD discriminator 0x03E800FF.

It is possible, although foreseen to be extremely rare, for identifiers of different BFD target identifier types to map to same BFD discriminator. How such conflict is to be resolved is outside the scope of this document.

5. BFD Target Identifier Table

Each network node is responsible for creating and maintaining a table that contains BFD discriminators, BFD target identifier types and BFD target identifiers. Intention of this table is to allow local entities to perform following lookups:

- BFD discriminator to BFD target identifier type and BFD target identifier
- BFD target identifier type and BFD target identifier to BFD discriminator

This table MUST contain entries for all locally reserved BFD discriminators and corresponding information. This table MAY need to contain entries from other network nodes, depending on the BFD target identifier type.

6. Reflector BFD Session

Each network node MUST create one or more reflector BFD sessions. This reflector BFD session is a session which transmits BFD control packets in response to received valid locally destined BFD control packets. Specifically, this reflector BFD session is to have following characteristics:

- Does not transmit any BFD control packets based on local timer expiry.
- Transmits BFD control packet in response to received valid locally destined BFD control packet.
- Capable of sending only two states: UP and ADMINDOWN.

One reflector BFD session can be responsible for handling response to received BFD control packets targeted to all local BFD target identifiers, or few reflector BFD sessions can each be responsible for subset of local BFD target identifiers. This policy is a local matter, and is outside the scope of this document.

In addition, a reflector BFD session MUST be address family agnostic. Single reflector BFD session MUST be able to handle incoming BFD control packets in IPv4 and IPv6, and MUST be able to respond with BFD control packets using same address family as received packets.

7. Full Reachability Validations

7.1. Initiator Behavior

Any network node can attempt to perform a full reachability validation to any BFD target identifier on other network nodes, as long as destination BFD target identifier is provisioned to use this mechanism. Transmitted BFD control packet by the initiator is to have "your discriminator" corresponding to destination BFD target identifier.

A node that initiates a BFD control packet can create an active BFD session to periodically send BFD control packet to a target, or BFD control packet can be crafted and sent out on "as needed basis" (ex: BFD ping) without any session presence. In both cases, a BFD instance MUST have unique "my discriminator" value assigned. If a node is to create multiple BFD instances to a same BFD target identifier, then each instance MUST have separate "my discriminator" values assigned.

If BFD control packet is to be sent via IP path, then:

- Destination IP address MUST be an IP address corresponding to target identifier.
- Source IP address MUST be a local IP address.
- IP TTL MUST be 255 for full reachability validations. Partial reachability validations MAY use smaller TTL value (see Section 8).
- One of well-known UDP destination ports for IP based BFD: 3784 for singlehop, 4784 for multihop, 6784 for BFD for LAG

If BFD control packet is to be sent via explicit label switching, then:

- BFD control packet MUST get imposed with a label stack that is expected to reach the target node.
- MPLS TTL MUST be 255 for full reachability validations. Partial reachability validations MAY use smaller TTL value (see Section 8).
- Destination IP address MUST be 127/8 for IPv4 and 0:0:0:0:0:FFFF:7F00/104 for IPv6.
- Source IP address MUST be a local IP address.
- IP TTL=1.
- Well-known UDP destination port for MPLS based BFD: 3784

7.2. Responder Behavior

A network node which receives BFD control packets transmitted by an initiator is referred as responder. Responder, upon reception of BFD control packets, is to perform necessary relevant validations described in [RFC5880]/[RFC5881]/[RFC5883]/[RFC5884]/[RFC5885].

7.2.1. Responder Demultiplexing

When responder receives a BFD control packet, if "your discriminator" value is not one of local entries in the BFD target identifier table, then this packet MUST NOT be considered for this mechanism. If "your discriminator" value is one of local entries in the BFD target identifier table, then the packet is determined to be handled by a reflector BFD session responsible for specified BFD targeted identifier. If the packet was determined to be processed further for this mechanism, then chosen reflector BFD session is to transmit a response BFD control packet using procedures described in Section 7.2.2, unless prohibited by local administrative or local policy reasons.

7.2.2. Reflector BFD Session Procedures

BFD target identifier type MUST be used to determine further information on how to reach back to the initiator.

In addition, destination IP address of received BFD control packet MUST be examined to determine how to construct response BFD control packet to send back to the initiator.

If destination IP address of received BFD control packet is not 127/8 for IPv4 or 0:0:0:0:0:FFFF:7F00/104 for IPv6, then:

- Destination IP address MUST be copied from received source IP address.
- Source IP address MUST be copied from received destination IP address if received destination IP address is a local address. Otherwise local IP address MUST be used.
- IP TTL MUST be 255.

If destination IP address of received BFD control packet is 127/8 for IPv4 or 0:0:0:0:0:FFFF:7F00/104 for IPv6, then received IP destination MUST be further examined to determine response transport options. If last 23 bits of 127/8 for IPv4 and 0:0:0:0:0:FFFF:7F00/104 for IPv6 is zero, then response SHOULD be label switched but MAY be IP routed. If last 23 bits of 127/8 for IPv4 and 0:0:0:0:0:FFFF:7F00/104 for IPv6 is not zero, then response SHOULD be label switched and SHOULD NOT be IP routed. Description of 23 bits is described in Section 9.

If BFD control packet response is determined to be IP routed, then:

- Destination IP address MUST be copied from received source IP address.
- Source IP address MUST be a local address.
- IP TTL MUST be 255.

If BFD control packet response is determined to be label switched, then:

- BFD control packet MUST get label switched back to the initiator. How label stack to be imposed on a response BFD control packet is determined for all cases is outside the scope of this document.
- MPLS TTL MUST be 255.
- Destination IP address MUST be 127/8 for IPv4 and 0:0:0:0:0:FFFF:7F00/104 for IPv6.
- Source IP address MUST be a local IP address.
- IP TTL MUST be 1.

Regardless of the response type, BFD control packet being sent by the responder MUST perform following procedures:

- Copy "my discriminator" from received "your discriminator", and "your discriminator" from received "my discriminator".

- UDP destination port MUST be same as received UDP destination port.

7.3. Further Packet Details

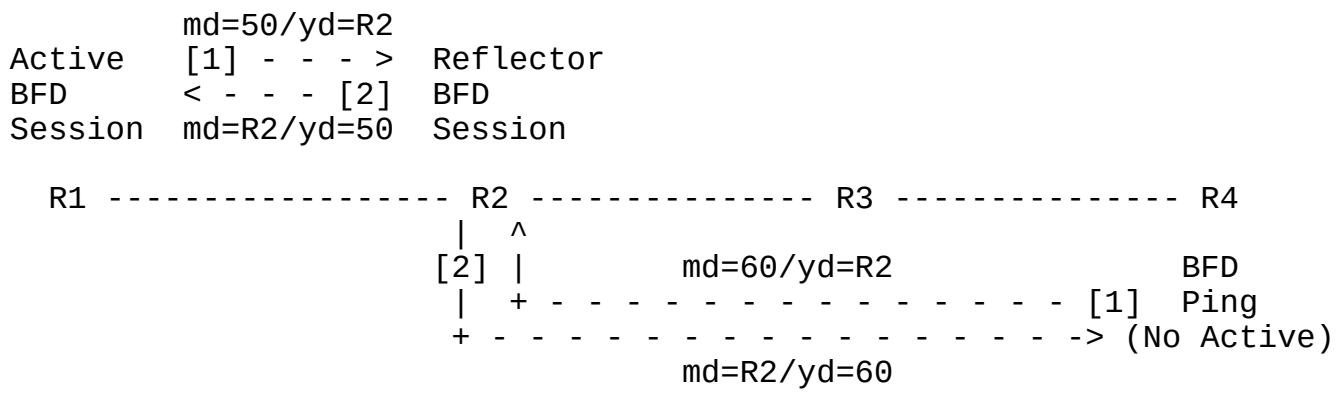
Further details of BFD control packets sent by initiator (ex: active BFD session):

- UDP destination port described in [RFC5881]/[RFC5883]/[RFC5884]/[RFC5885].
- UDP source port as per described in [RFC5881]/[RFC5883]/[RFC5884]/[RFC5885].
- "my discriminator" assigned by local node.
- "your discriminator" corresponding to an identifier of target node.
- "State" MUST be set to a value reflecting local state.
- "Desired Min TX Interval" MUST be set to a value reflecting local desired minimum transmit interval.
- "Required Min RX Interval" MUST be zero.
- "Required Min Echo RX Interval" SHOULD be zero.
- "Detection Multiplier" MUST be set to a value reflecting locally used multiplier value.

Further details of BFD control packets sent by responder (reflector BFD session):

- UDP destination port described in [RFC5881]/[RFC5883]/[RFC5884]/[RFC5885].
- UDP source port as per described in [RFC5881]/[RFC5883]/[RFC5884]/[RFC5885].
- "my discriminator" MUST be copied from received "your discriminator".
- "your discriminator" MUST be copied from received "my discriminator".
- "State" MUST be UP or ADMINDOWN. Usage of ADMINDOWN state is described in Section 7.6.
- "Desired Min TX Interval" MUST be copied from received "Desired Min TX Interval".
- "Required Min RX Interval" MUST be set to a value reflecting how many incoming control packets this reflector BFD session can handle.
- "Required Min Echo RX Interval" SHOULD be set to zero.
- "Detection Multiplier" MUST be copied from received "Detection Multiplier".

Simple ASCII art is provided to illustrate the concept described so far.



7.4. Diagnostic Values

Diagnostic value in both directions MAY be set to a certain value, to attempt to communicate further information to both ends. However, details of such are outside the scope of this specification.

7.5. Additional Initiator Behavior

- If initiator receives valid BFD control packet in response to transmitted BFD control packet, then initiator SHOULD conclude that packet reached intended target.

- How many repeated absence of response should make initiator consider loss of reachability, and what action would be triggered as result are outside the scope of this specification.

7.6. Additional Responder Behavior

- BFD control packets transmitted by a reflector BFD session MUST have "Required Min RX Interval" set to a value which reflects how many incoming control packets this reflector BFD session can handle. Responder can control how fast initiators will be sending BFD control packets to self by ensuring "Required Min RX Interval" reflects a value based on current load.

- If a reflector BFD session wishes to communicate to some or all initiators that monitored BFD target identifier is "temporary out of service", then BFD control packets with "state" set to ADMINDOWN are sent to those initiators. Initiators, upon reception of such packets, MUST NOT conclude loss of reachability to corresponding BFD target identifier, and MUST back off packet transmission interval to corresponding BFD target identifier an interval no faster than 1 second.

8. Partial Reachability Validations

Same mechanism as described in "Full Reachability Validations" section will be applied with exception of following differences on initiator.

- When initiator wishes to perform a partial reachability validation towards identifier X on identifier Y, number of hops to identifier Y is calculated.

- TTL value based on this calculation is used as the IP TTL or MPLS TTL on top most label, and "your discriminator" of transmitted BFD control packet will carry BFD discriminator corresponding to target transit identifier Y.

- Imposed label stack or IP destination address will continue to be of identifier X.

9. MPLS Label Verifications

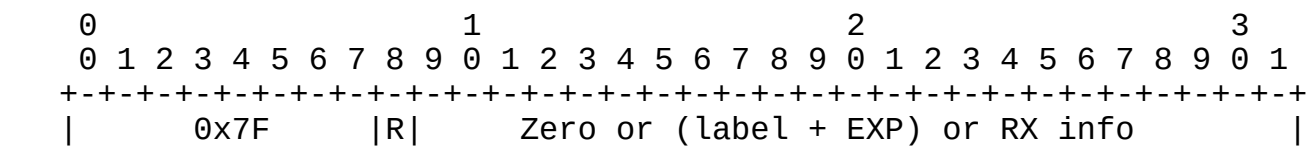
This section is only applicable to MPLS based sessions using this mechanism.

9.1. MPLS Label Verifications Mechanism

With full and partial reachability validations, initiator has the ability to determine if target identifier received the packet on any interfaces. This section describes additional mechanism for initiator to determine if target identifier received the packet on a specific interface.

So far for MPLS based sessions, this mechanism makes use of destination IP address of 127/8 range for IPv4 and of 0:0:0:0:0:FFFF:7F00/104 range for IPv6, in both directions. In this section, 127/8 will be used to describe the MPLS label verification mechanism. However, same concept is to be applied to IPv6 range 0:0:0:0:0:FFFF:7F00/104.

When a network node wishes to perform MPLS label verification, BFD control packet will have lower 23 bits of 127/8 destination IP address embedded with (label value + EXP) that is used to reach intended target identifier. Receiver of this BFD control packet, if last 23 bits of 127/8 address is not zero, then will embed information reflecting how the packet was received in the lower 23 bits of 127/8 destination IP address in the response BFD control packet. If responder received the BFD control packet on a non-point-to-point interface, source MAC address MAY need to be examined to determine the "RX info" to embed in the returning packet.



9th bit is reserved for the time being and SHOULD be set to zero and SHOULD be ignored on receipt, by both initiator and responder

Initiator receiving back a response will know that packet did reach intended identifier. Initiator can also look into lower 23 bits of IP destination address in received BFD control packet to determine if packet sent was received by intended identifier in expected way (ex: expected RX interface).

When (label + EXP) is being encoded, label is specified in higher 20 bits of 23 bits and EXP is specified in lower 3 bits of 23 bits.

If a response BFD control packet is received, then initiator can conclude that a packet has reached intended node correctly. With information embedded in last 23 bits of response BFD control packet from responder, initiator has the ability to perform further verifications on how responded node received BFD control packet.

9.2. Localhost Address Usage

Last 23 bits of 127/8 for IPv4 and 0:0:0:0:0:FFFF:7F00/104 for IPv6 being non-zero is the trigger for responder to embed RX information in the response. When initiator is performing only reachability validations to target identifiers, then last 23 bits of the localhost address MUST be zero. This is to ensure unnecessary processing at responder is eliminated.

10. Scaling Aspect

This mechanism brings forth one noticeable difference in terms of scaling aspect: number of BFD sessions. This specification eliminates the need for egress nodes to have fully active BFD sessions when only side is desired to perform reachability validations. With introduction of reflector BFD concept, egress no longer is required to create any active BFD session per path/LSP basis. Due to this, total number of BFD sessions in a network is reduced.

If traditional BFD technology was used on a network comprised of N nodes, and each node monitored M uni-directional paths/LSPs, then total number of BFD sessions in such network will be:

(((N - 1) x M) x 2)

Assuming that each network node creates one reflector BFD session to handle all local BFD target identifiers, then total number of BFD sessions in same scenario will be:

(((N - 1) x M) + N)

11. Co-existence with Traditional BFD

This mechanism has no issues being deployed with traditional BFDs ([RFC5881]/[RFC5883]/[RFC5884]/[RFC5885]) because BFD discriminators which allow this mechanism to function are explicitly reserved.

12. BFD Echo

BFD echo is outside the scope of this document.

13. Security Considerations

Same security considerations as [RFC5880], [RFC5881], [RFC5883], [RFC5884] and [RFC5885] apply to this document.

Additionally, implementing following measures will strengthen security aspects of this mechanism described by this document.

- Implementations MUST provide filtering capability based on source IP addresses or source node segment IDs of received BFD control packets: [RFC2827].

- Implementations MUST NOT act on received BFD control packets containing Martian addresses as source IP addresses.

- Implementations MUST ensure response target IP addresses or node segment IDs are reachable.

14. IANA Considerations

BFD Target Identifier types:

Value	BFD Target Identifier Type
0	Reserved
1	IP (IPv4 Address and Router ID)
2	Segment Routing Node Segment ID

15. Acknowledgements

Authors would like to thank Marc Binderberger from Cisco Systems for providing valuable comments.

16. Contributing Authors

Tarek Saad
Cisco Systems
Email: tsaad@cisco.com

Siva Sivabalan
Cisco Systems
Email: msiva@cisco.com

Nagendra Kumar
Cisco Systems
Email: naikumar@cisco.com

17. References

17.1. Normative References

[I-D.previdi-filsfils-isis-segment-routing] Previdi, S., Filsfils, C., Bashandy, A., Horneffer, M., Decraene, B., Litkowski, S., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., and J. Tantsura, "Segment Routing with IS-IS Routing Protocol", draft-previdi-filsfils-isis-segment-routing-02 (work in progress), March 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

[RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, June 2010.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

17.2. Informative References

[I-D.ietf-bfd-on-lags]
Bhatia, M., Chen, M., Boutros, S., Binderberger, M., and J. Haas, "Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces", draft-ietf-bfd-on-lags-00 (work in progress), May 2013.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

[RFC5885] Nadeau, T. and C. Pignataro, "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.

[RFC6428] Allan, D., Swallow Ed. , G., and J. Drake Ed. , "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, November 2011.

Authors' Addresses

Nobo Akiya
Cisco Systems

Email: nobo@cisco.com

Carlos Pignataro
Cisco Systems

Email: cpignata@cisco.com

Dave Ward
Cisco Systems

Email: wardd@cisco.com