Internet Engineering Task Force (IETF)
Request for Comments: 7130
Category: Standards Track
ISSN: 2070-1721

M. Bhatia, Ed.
Alcatel-Lucent
M. Chen, Ed.
Huawei Technologies
S. Boutros, Ed.
M. Binderberger, Ed.
Cisco Systems
J. Haas, Ed.
Juniper Networks

February 2014

Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces

Abstract

This document defines a mechanism to run Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) interfaces. It does so by running an independent Asynchronous mode BFD session on every LAG member link.

This mechanism allows the verification of member link continuity, either in combination with, or in absence of, Link Aggregation Control Protocol (LACP). It provides a shorter detection time than what LACP offers. The continuity check can also cover elements of Layer 3 (L3) bidirectional forwarding.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7130.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

The Bidirectional Forwarding Detection (BFD) protocol [RFC5880] provides a mechanism to detect faults in the bidirectional path between two forwarding engines, including interfaces, data links, and to the extent possible the forwarding engines themselves, with potentially very low latency. The BFD protocol also provides a fast mechanism for detecting communication failures on any data links and the protocol can run over any media and at any protocol layer.

LAG, as defined in [IEEE802.1AX], provides mechanisms to combine multiple physical links into a single logical link. This logical link provides higher bandwidth and better resiliency, because if one

of the physical member links fails, the aggregate logical link can continue to forward traffic over the remaining operational physical member links.

Currently, the Link Aggregation Control Protocol (LACP) is used to detect failures on a per-physical-member link. However, the use of BFD for failure detection would (1) provide a faster detection, (2) provide detection in the absence of LACP, and (3) would be able to verify the ability for each member link to be able to forward L3 packets.

Running a single BFD session over the aggregation without internal knowledge of the member links would make it impossible for BFD to guarantee detection of the physical member link failures.

The goal is to verify link Continuity for every member link. This corresponds to [RFC5882], Section 7.3.

The approach taken in this document is to run an Asynchronous mode BFD session over each LAG member link and make BFD control whether the LAG member link should be part of the L2 load-balancing table of the LAG interface in the presence or the absence of LACP.

This document describes how to establish an Asynchronous mode BFD session per physical LAG member link of the LAG interface.

While there are native Ethernet mechanisms to detect failures (802.1ax, .3ah) that could be used for LAG, the solution defined in this document enables operators who have already deployed BFD over different technologies (e.g., IP, MPLS) to use a common failure detection mechanism.

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. BFD on LAG Member Links

The mechanism defined for a fast detection of LAG member link failure is to run Asynchronous mode BFD sessions on every LAG member link. We call these per-LAG-member-link BFD sessions "micro-BFD sessions" in the remainder of this document.

2.1. Micro-BFD Session Address Family

Member link micro-BFD sessions, when using IP/UDP encapsulation, can use IPv4 or IPv6 addresses. Two micro-BFD sessions MAY exist per member link: one IPv4 another IPv6. When an address family is used on one member link, then it MUST be used on all member links of the particular LAG.

2.2. Micro-BFD Session Negotiation

A single micro-BFD session for every enabled address family runs on each member link of the LAG. The micro-BFD session's negotiation MUST follow the same procedures defined in [RFC5880] and [RFC5881].

Only Asynchronous mode BFD is considered in this document; the use of the BFD echo function is outside the scope of this document. At least one system MUST take the Active role (possibly both). The micro-BFD sessions on the member links are independent BFD sessions. They use their own unique local discriminator values, maintain their own set of state variables, and have their own independent state machines. Timer values MAY be different, even among the micro-BFD sessions belonging to the same aggregation, although it is expected that micro-BFD sessions belonging to the same aggregation will use the same timer values.

The demultiplexing of a received BFD packet is solely based on the Your Discriminator field, if this field is nonzero. For the initial Down BFD packets of a BFD session, this value MAY be zero. In this case, demultiplexing MUST be based on some combination of other fields that MUST include the interface information of the member link and the destination UDP port of the received BFD packet.

The procedure for the reception of BFD control packets in Section 6.8.6 of [RFC5880] is amended as follows for per-LAG-member-link micro-BFD sessions:

> If the Your Discriminator field is nonzero and a micro-BFD over a LAG session is found, the interface on which the micro-BFD control packet arrived MUST correspond to the interface associated with that session.

This document defines the BFD control packets for each micro BFD session to be IP/UDP encapsulated as defined in [RFC5881], but with a new UDP destination port 6784.

The new UDP port removes the ambiguity of BFD over LAG packets from BFD over single-hop IP. An example is (mis-)configuring a LAG with micro-BFD sessions on one side but using a [RFC5881] BFD session for the LAG (treated as a single interface) on the opposite side.

The procedures in this document MUST be used for BFD messages addressed to port 6784 and MUST NOT be used for others ports assigned in RFCs describing other BFD modes.

Control packets use a destination IP address that is configured on the peer system and can be reached via the LAG interface.

Implementations may range from explicitly configuring IP addresses for the BFD sessions to out-of-band methods for learning the destination IP address. The details are outside the scope of this document.

2.3. Micro-BFD Session Ethernet Details

On Ethernet-based LAG member links, the destination Media Access Control (MAC) is the dedicated multicast MAC address 01-00-5E-90-00-01 to be the immediate next hop. This dedicated MAC address MUST be used for the initial BFD packets of a micro-BFD session when in the Down/AdminDown and Init states. When a micro-BFD session is changing into the Up state, the first bfd.DetectMult packets in the Up state MUST be sent with the dedicated MAC. For BFD packets in the Up state following the first bfd.DetectMult packets, the source MAC address from the received BFD packets for the session MAY be used instead of the dedicated MAC.

All implementations MUST be able to send and receive BFD packets in Up state using the dedicated MAC address. Implementations supporting both, sending BFD Up packets with the dedicated and the received MAC, need to offer means to control the behaviour.

On Ethernet-based LAG member links, the source MAC SHOULD be the MAC address of the member link transmitting the packet.

This mechanism helps to reduce the use of additional MAC addresses, which reduces the required resources on the Ethernet hardware on the receiving member link.

Micro-BFD packets SHOULD always be sent untagged. However, when the LAG is operating in the context of IEEE 802.1q or IEEE 802.qinq, the micro-BFD packets may either be untagged or be sent with a vlan tag of Zero (802.1p priority tagged). Implementations compliant with this standard MUST be able to receive both untagged and 802.1p priority tagged micro-BFD packets.

3. Interaction between LAG and BFD

The micro-BFD sessions for a particular LAG member link MUST be requested when a member link state is either Distributing or Standby. The sessions MUST be deleted when the member link is in neither Distributing nor Standby state anymore.

BFD is used to control if the load-balancing algorithm is able to select a particular LAG member link. In other words, even when Link Aggregation Control Protocol (LACP) is used and considers the member link to be ready to forward traffic, the member link MUST NOT be used by the load balancer until all the micro-BFD sessions of the particular member link are in Up state.

In case an implementation has separate load-balancing tables for IPv4 and IPv6 and if both an IPv4 and IPv6 micro-BFD session exist for a member link, then an implementation MAY enable the member link in the load-balancing algorithm based on the BFD session with a matching address family alone.

An exception is the BFD packet itself. Implementations MAY receive and transmit BFD packets via the Aggregator’s MAC service interface, independent of the session state.

4. BFD on LAG Member Links and L3 Applications

The mechanism described in this document is likely to be used by modules managing Interfaces or LAGs and, thus, managing the member links of a LAG. Typical L3 protocols like OSPF do not have an insight into the LAG and treat it as one bigger interface. The signaling from micro sessions to L3 protocols is effectively done by the impact of micro-BFD sessions on the load-balancing table and the Interface/LAG managing module’s potential decision to shut down the LAG. An active method to test the impact of micro-BFD sessions is for L3 protocols to request a single BFD session per LAG.

5. Detecting a Member Link Failure

When a micro-BFD session goes down, this member link MUST be taken out of the LAG load-balancing table(s).

In case an implementation has separate load-balancing tables for IPv4 and IPv6, then if both an IPv4 and IPv6 micro-BFD session exist for a member link, an implementation MAY remove the member link only from the load-balancing table that matches the address family of the failing BFD session. For example, the IPv4 micro-BFD session fails but the IPv6 micro-BFD session stays Up, then the member link MAY be removed from only the IPv4 load balance table; the link MAY remain in

the IPv6 load-balancing table. Alternatively, the member link may be removed from both the IPv4 and IPv6 load-balancing tables. This decision is an implementation detail.

6. Security Considerations

This document does not introduce any additional security issues and the security mechanisms defined in [RFC5880] apply in this document.

7. IANA Considerations

IANA assigned a dedicated MAC address 01-00-5E-90-00-01 (see [RFC7042]) as well as UDP port 6784 for Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces. IANA has changed the reference to [RFC7130].

IANA has changed the registry for port 6784 to show the Assignee as [IESG] and the Contact as [BFD_Chairs]. The expansion of [BFD_Chairs] is shown as "mailto:bfd-chairs@tools.ietf.org". IANA has changed the reference to [RFC7130].

8. Acknowledgements

We would like to thank Dave Katz, Alexander Vainshtein, Greg Mirsky, and Jeff Tantsura for their comments.

The initial event to start the current discussion was the distribution of "Bidirectional Forwarding Detection (BFD) for Interface" (July 2011).

9. Contributors

Paul Hitchen
BT
EMail: paul.hitchen@bt.com

George Swallow
Cisco Systems
EMail: swallow@cisco.com

Wim Henderickx
Alcatel-Lucent
EMail: wim.henderickx@alcatel-lucent.com

Nobo Akiya
Cisco Systems
EMail: nobo@cisco.com

Neil Ketley
Cisco Systems
EMail: nketley@cisco.com

Carlos Pignataro
Cisco Systems
EMail: cpignata@cisco.com

Nitin Bahadur
Bracket Computing
EMail: nitin@brkt.com

Zuliang Wang
Huawei Technologies
EMail: liang_tsing@huawei.com

Liang Guo
China Telecom
EMail: guoliang@gsta.com

Jeff Tantsura
Ericsson
EMail: jeff.tantsura@ericsson.com

10. References

10.1. Normative References

[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.

[RFC5882] Katz, D. and D. Ward, "Generic Application of Bidirectional Forwarding Detection (BFD)", RFC 5882, June 2010.

10.2. Informative References

[IEEE802.1AX] IEEE Std. 802.1AX, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", November 2008.

[RFC7042] Eastlake, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, October 2013.

Appendix A. Considerations When Using BFD on Member Links

If the BFD-over-LAG feature were provisioned on an aggregated link member after the link was already active within a LAG, BFD session state should not influence the load-balancing algorithm until the BFD session state transitions to Up. If the BFD session never transitions to Up but the LAG becomes inactive, the previously documented procedures would then normally apply.

This procedure ensures that the sequence of events -- enabling the LAG and enabling BFD on the LAG -- has no impact on the forwarding service.

If the BFD-over-LAG feature were deprovisioned on an aggregate link member while the associated micro-BFD session was in Up state, BFD should transition its state to AdminDown and should attempt to communicate this state change to the peer.

If the local or the remote state of a micro-BFD session is AdminDown, the system should not indicate a connectivity failure to any client and should not remove the particular LAG member link from forwarding. This behaviour is independent from the use of Link Aggregation Control Protocol (LACP) for the LAG.

When traffic is forwarded across a link while the corresponding micro-BFD session is not in Up state, an implementation may use a configurable timeout value after which the BFD session must have reached Up state otherwise the link is taken out of forwarding.

When such timeout values exist, the configuration must allow the ability to turn off the timeout function.

The configurable timeout value shall ensure that a LAG is not remaining forever in an "inconsistent" state where forwarding occurs on a link with no confirmation from the micro-BFD session that the link is healthy.

Note that if one device is not operating a micro-BFD session on a link, while the other device is and perceives the session to be Down, this will result in the two devices having a different view of the status of the link. This would likely lead to traffic loss across the LAG. The use of another protocol to bootstrap BFD can detect such mismatched config, since the side that’s not configured can send a rejection error. Such bootstrapping mechanisms are outside the scope of this document.

Authors' Addresses

Manav Bhatia (editor)
Alcatel-Lucent
Bangalore 560045
India

EMail: manav.bhatia@alcatel-lucent.com

Mach(Guoyi) Chen (editor)
Huawei Technologies
Q14 Huawei Campus, No. 156 Beiqing Road, Hai-dian District
Beijing 100095
China

EMail: mach@huawei.com

Sami Boutros (editor)
Cisco Systems

EMail: sboutros@cisco.com

Marc Binderberger (editor)
Cisco Systems

EMail: mbinderb@cisco.com

Jeffrey Haas (editor)
Juniper Networks

EMail: jhaas@juniper.net