Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 13, 2013

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

April 11, 2013

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

draft-mmm-bfd-on-lags-07

Abstract

This document proposes a mechanism to run 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, LACP. It provides a shorter detection time than what LACP offers. The continuity check can also cover elements of layer 3 bidirectional forwarding.

This mechanism utilizes a well-known UDP port distinct from that of single-hop BFD over IP. This new UDP port removes the ambiguity of BFD over LAG packets from BFD over single-hop IP.

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 October 13, 2013.

Copyright Notice

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 link(s), 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.

Link aggregation (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 since 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 (3) and would be able to verify L3 Continuity per member link.

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 a Asynchronous mode BFD session over each member link and make BFD control whether the member link should be part of the L2 Loadbalance table of the LAG virtual port in the presence or the absence of LACP.

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

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

2. BFD on LAG member links

The mechanism proposed 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 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 which MUST include the interface information of the member link.

The procedure for the Reception of BFD Control Packets in Section 6.8.6 of [RFC5880] is amended as follows for per member link micro BFD over LAG sessions: "If the Your Discriminator field is non-zero and a micro BFD over LAG session is found, the interface on which the micro BFD control packet arrived on 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.

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

details of how this destination IP address is learned are outside the scope of this document.

2.3. Micro BFD session Ethernet details

On Ethernet-based LAG member links the destination 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 state. When a micro BFD session is changing into Up state then the first bfd.DetectMult packets with Up state MUST be sent with the dedicated MAC. For the following BFD packets with Up state the 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 port 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 port.

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 sent with a vlan tag of Zero (802.1p priority tagged). Implementations compliant to this standard MUST be able to receive both untagged and 802.1p priority tagged micro BFD packets.

3. LAG Management Module

3.1. Interaction between LAG and BFD

The LAG Management Module (LMM) could be envisaged as a client of BFD; i.e. the LMM requests the micro BFD sessions per member link. The LMM then uses the micro BFD session state, in addition to LACP state, to monitor the health of the individual members links of the LAG.

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

in Standby state anymore.

BFD is used to control if the load balance algorithm is able to select a particular port. In other words, even when LACP is used and considers the member link to be ready to forward traffic, the member link is only used by the load balancer when all the micro BFD sessions of the member link are Up.

In case an implementation has separate load balance tables for IPv4 and IPv6 then if both an IPv4 and IPv6 micro session exist for a member link an implementation MAY enable the member link in the distribution algorithm only when the BFD session with a matching address family is changing into Up state.

An exception are the BFD packets itself. Implementations MAY receive and transmit BFD packets via the Aggregator's MAC service interface independent of the session state.

3.2. Handling Exceptions

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 balance 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.

If the BFD over LAG feature were deprovisioned on an aggregate link member after the BFD session had transitioned to Up, BFD MAY indicate to the remote port that it should not take the port down or remove it from the aggregation by setting its BFD session state to AdminDown.

When a micro BFD session receives AdminDown from the peer, it is RECOMMENDED to have a configurable timeout value. If the BFD session has not been removed within the timeout period the link is taken out of forwarding.

When traffic is forwarded across a link before the corresponding micro BFD session is Up it is RECOMMENDED to have a configurable timeout value after which the BFD session must have reached Up state or otherwise the link is taken out of forwarding.

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.

4. BFD on LAG members and layer-3 applications

The mechanism described in this document is likely to be used by modules like LMM or some Interface management module. Typical layer 3 protocols like OSPF do not have an insight into the LAG and treat it as one bigger interface. The signalling from micro sessions to layer 3 protocols is effectively done by the impact of BFD micro sessions on the load balance table and the LMM's potential decision to shut down the LAG. An active method to test the impact of micro sessions is for layer 3 protocols to request a single BFD session per LAG.

5. Detecting a member link failure

When a micro BFD session goes down then this member link MUST be taken out of the LAG L2 load balance table(s).

In case an implementation has separate load balance tables for IPv4 and IPv6 then if both an IPv4 and IPv6 micro session exist for a member link an implementation MAY remove the member link from the load balance table only that matches the address family of the failing BFD session. If for example the IPv4 micro session fails but the IPv6 micro session stays up then the member link MAY be removed from the IPv4 load balance table only but remains forwarding in the IPv6 load balance table.

6. Security Consideration

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 as well as UDP port 6784 for UDP encapsulated micro BFD sessions.

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 draft-chen-bfd-interface-00.

9. Contributing authors

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
Juniper Networks
Email: nitinb@juniper.net

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.

[RFC5342] Eastlake, D., "IANA Considerations and IETF Protocol Usage for IEEE 802 Parameters", BCP 141, RFC 5342, September 2008.

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