Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 3, 2012

M. Bhatia, Ed.
Alcatel-Lucent
M. Chen, Ed.
Huawei Technologies
S. Boutros, Ed.
M. Binderberger, Ed.
Cisco Systems
J. Haas, Ed.
Juniper Networks
June 1, 2012

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

draft-mmm-bfd-on-lags-04

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 connectivity, either in combination with, or in absence of, LACP. It provides a shorter detection time than what LACP offers. The connectivity 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

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 connectivity 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 connectivity for every member link.

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

Only one address family MUST be used for all micro BFD sessions running on all LAG member links. I.e. all member link micro BFD sessions MUST either use IPv4 or IPv6.

2.2. Micro BFD session negotiation

A single micro BFD session 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."

The BFD Control packets for each micro BFD session are IP/UDP encapsulated as defined in [RFC5881], but with a new UDP destination port "BfdBndlPort" (to be assigned by IANA). Control packets use a destination IP address that is the peer's remote IP address. The details of how this destination IP address is learned is outside the scope of this document.

On Ethernet-based LAG member links the destination MAC is a dedicated MAC address (to be assigned by IANA according to [RFC5342]) 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 state. When sending BFD packets for the micro BFD session in the Init and Up state, the MAC address from the received BFD packets for the session MUST be used.

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.

3. LAG Management Module

The LAG Management Module (LMM) could be envisaged as a client of BFD; i.e. the LMM requests a micro BFD session 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 session for a particular port MUST be requested when the port is attached to an aggregator. The session MUST be deleted when the port is detached from the aggregator.

The LMM uses the status of the BFD session to determine whether the member link should be included in the LAG L2 load balance table. 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 traffic load balancer when the micro BFD session is Up.

3.1. BFD in the presence of LACP

Prior to LACP coming up, the micro BFD session is Passive and does not send BFD control packets. Once LACP has determined that a link is suitable for aggregation within its selected LAG and has completed negotiations with the partner device so as to bring that link to Distributing state, the micro BFD session can be made Active and the session started on the link.

BFD, as a layer 3 protocol, is viewed as running across the LAG, with load balancing constraints ensuring particular BFD micro sessions are effectively bound to particular member links.

Although the link is in LACP Distributing state, it should not be used for carrying traffic other than the micro BFD session. BFD is used to verify that a link is ready to be an active member of its LAG for the purpose of carrying LAG level data traffic. Only when the

micro BFD session is up should the link become active for forwarding general traffic over the bundle.

3.2. BFD in the absence of LACP

Use of LACP for link aggregation is strictly optional. It is equally possible to use no aggregation control protocol and to step directly from the layer 1 or layer 2 OAM state becoming operational to starting the micro BFD session. In this case, the micro BFD sessions begin as Active.

3.3. 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 link state 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.

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

Layer 3 protocols, e.g. OSPF, may use a micro BFD session to detect failures on the LAG virtual port, or may establish a new BFD session over the logical LAG virtual port.

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.

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

The IANA is requested to assign a well-known port number for the UDP encapsulated micro BFD sessions. IANA is also requested to assign a dedicated MAC address according to RFC 5342 [RFC5342].

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.

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