INTERNET-DRAFT
Intended Status: Historic
Expires: May 20, 2012

A.Palanivelan
EMC Corporation
Nov 17, 2011

A Record of Discussions of Graceful Restart Extensions for Bidirectional Forwarding Detection (BFD) draft-palanivelan-bfd-v2-gr-13

Status of this Memo

Copyright and License Notice

Abstract

This document is a historical record of discussions about extending the Bidirectional Forwarding Detection (BFD) protocol to provide additional capabilities to handle Graceful Restart.

These discussions took place in the context of the IETF's BFD working group, and the consensus in that group was that these extensions should not be made.

This document presents a summary of the challenges to BFD in surviving a graceful restart, and outlines a potential protocol solution that was discussed and rejected within the BFD working group. The purpose of this document is to provide a record of the work done so that future effort will not be wasted. This document does not propose or document any extensions to BFD, and there is no intention that it should be implemented in its current form.

Table of Contents

1 Introduction

The Bidirectional Forwarding Detection protocol [BFD] provides mechanism for liveness detection of arbitrary paths between systems. It is intended to provide low-overhead, short-duration detection of failures in the path between adjacent forwarding engines, including the interfaces, data links, and to the extent possible, the forwarding engines themselves. It operates independently of media, data protocols, and routing protocols. An additional BFD goal is to provide a single mechanism that can be used for liveness detection over any media, at any protocol layer, with a wide range of detection times and overhead, to avoid a proliferation of different methods.

Graceful Restart (GR) was considered for BFD, but was rejected by the BFD Working Group as unnecessary and potentially detrimental to the protocol. As a result, the work on BFD GR was not progressed within the working group.

This document presents a summary of the challenges to BFD in surviving a graceful restart, and outlines a potential protocol solution that was discussed and rejected within the BFD working group. The purpose of this document is to provide a record of the work done so that future effort will not be wasted.

This document does not propose or document any extensions to BFD, and there is no intention that it should be implemented in its current form.

2 Overview

The Bidirectional Forwarding Detection [BFD] specification defines a protocol with simple and specific semantics. Its purpose is to verify liveliness of the connectivity between a pair of systems, for a particular data protocol across a path (which may be of any technology, length, or at any protocol layer). The promptness of the detection of a path failure can be controlled by trading off protocol overhead and system load with detection times.

BFD works properly without a need for any GR support. The author is aware of BFD implementations that have experienced problems maintaining liveliness verification during GR. It is true that prioritizing BFD would make sure the other CPU intensive processes do not cause BFD to fail, but this may not be possible in some implementations as there may be other higher priority processing that can't be ignored. For example, perhaps existing subscriber connections can't be given a lesser priority.

3 Motivations

This section of the document discusses the following issues, might be seen in BFD deployments:

* Restarts with control protocols

* BFD co-existing with Broadband configurations

The later sections of this document capture ideas about protocol extensions that attempted to provide a general GR mechanism for BFD, but were rejected by the working group as unnecessary and inappropriate. The working group believes that the existing BFD design has the capabilities to take care of these challenges.

3.1 Restarts with control protocols

Section 4.3 of [RFC5882] describes how BFD can interact with the GR of control plane protocols.

Some protocols can signal the intention to perform a restart before initiating the restart, and can indicate when the restart has been completed. In these cases, [RFC5882] recommends that the restart should not be aborted and no topology change should be signalled in the control plane if BFD detects a session failure during the restart.

Control protocols that cannot signal a planned restart must treat planned and unplanned restarts in the same way. In order to avoid treating a BFD failure being triggered by the restart and causing the restart to be aborted or the topology modified, such protocols depend on the restarting system signalling the existence of the restart event before BFD detects the failure (for example, before the BFD session times out).

In most cases, whether the restart is planned or unplanned, it is likely that the BFD session timeout will be shorter that the restart time up to the point where the Graceful Restart event can be signalled. Thus, if there is an interruption to BFD caused by the restart, BFD will detect a fault and cause a topology change to be signaled. That means there could be an issue in implementations where a control plane restart event causes BFD to be disrupted. Such a situation could impact non-stop routing and non-stop forwarding support using GR-enabled protocols.

In considering this situation, the BFD working group determined that the solution was to implement the system such that a protocol restart

would not cause disruption to the function of the BFD session. Such could easily be achieved by maintaining the BFD session in the forwarding hardware. In arriving at this determination, the working group realized that any restart event that disrupted the BFD processes in the forwarding hardware would also result in a disruption to forwarding and it would, therefore, be correct to allow BFD to report the failure.

3.2 BFD Co-existing with Broadband configurations

Assume a Provider-Edge (PE) router may have active DHCP sessions with a large number of clients (say 16k). During a planned restart of the PE, it is possible that a number of the DHCP clients will request the server (restarting router) to renew client IP addresses. These requests will be retried and will reach the router in bulk after it has just come up. The router might be implemented to treat them at a high priority and respond to them. When there are thousands of such requests to the restarting router, the router might spend a major part of its first second of up time addressing them.

In this scenario, a control protocol like OSPFv2 that has GR enabled [OSPF-GRACE], could withstand the restart for the specified restart interval (as it will be in seconds) and it is likely to survive the restart, maintaining its forwarding plane.

In the same scenario, if BFD is enabled for OSPFv2 for an unplanned restart, the (BFD) neighbor router will be expecting BFD control packets in a milliseconds interval; during the restart process it could timeout, which would also impact the associated OSPFv2 adjacency and result in loss of traffic.

The scenario will be the same for BFD with a protocol such as IS-IS [IS-IS-GRACE], where the problem would be seen even for a planned restart since there is no ability to signal the restart event.

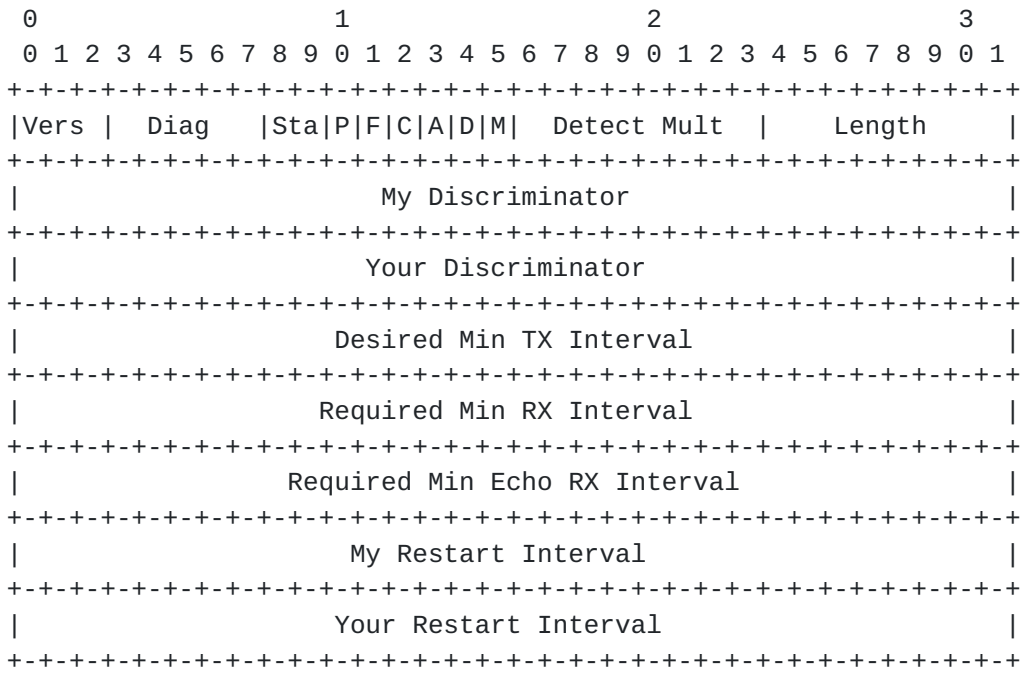
In its consideration of this scenario, the BFD working group decided that the situation would only arise if the PE router implementation gave disproportionate resources to other processes in such a way as to impact BFD. The working group considered that a sensible implementation would not lead to the problems described in this section.

4. Extensions to BFD

The protocol elements described in this section were rejected by the BFD working group and should not be implemented. The protocol procedures are not complete and the message formats do not form part of any BFD specification.

In discussions of a potential BFD protocol solution to circumvent the issues described in Section 3, a proposal was made to introduce a new BFD diag value to indicate that the neighbor is restarting, and provisions to configure BFD Graceful Restart timers.

The Generic BFD Control Packet [BFD] format shown below includes two additional fields "My Restart Interval" and "Your Restart Interval".



4.1 Diagnostic (Diag)

A diagnostic code specifying the local system's reason for the last change in session state. This field allows remote systems to determine the reason that the previous session failed. Values for this field are allocated by IANA according to Expert Review [RFC5226]. The procedures discussed, rejected and recorded in this document include a new diag value to represent "Neighbor Restarting".

4.2 State (Sta)

The procedures discussed, rejected and recorded in this document include a new BFD session state, "NeighborRestart".

The current session state is signalled in the Sta field [BFD]. This is a two-bit field and [BFD] defines all four possible values for the field. No mechanism was discussed or agreed upon within the BFD working group for how this additional session state would be signalled. A proposal to use the value 4 in the two-bit field was considered impractical.

4.3 My Restart Interval

The procedures discussed, rejected and recorded in this document include the addition of a My Restart Interval field to the BFD message as shown in the figure in Section 4.

This is the restart interval, in microseconds, of the transmitting system advertised to the remote system. In the case of a restart (of the transmitting system), the remote system is expected to keep the BFD session up for this duration of time. This field needs to have a value greater than the detection time (see section 6.2). A value of 0 indicates to the remote system that this system has BFD-GR disabled. The Length (L) field in the BFD header that indicates the fixed header length, would include the length of this field if this field were present.

4.4 Your Restart Interval

The proposed procedures discussed, rejected and recorded in this document include the addition of a Your Restart Interval field to the BFD message as shown in the figure in Section 4.

This is the restart interval, in microseconds, received from the corresponding remote system. In the case of a restart (of the remote system), the transmitting system is expected to keep the BFD session up for this duration of time. This field needs to have a value greater than the detection time. The Length (L) field in the BFD header that indicates the fixed header length, would include the length of this field, if this field were present.

5. State Machine for BFD with GR Support

The BFD state machine is quite straightforward and explained in detail by [BFD]. [BFD] describes different states for BFD as: Down, Init, Up, AdminDown.

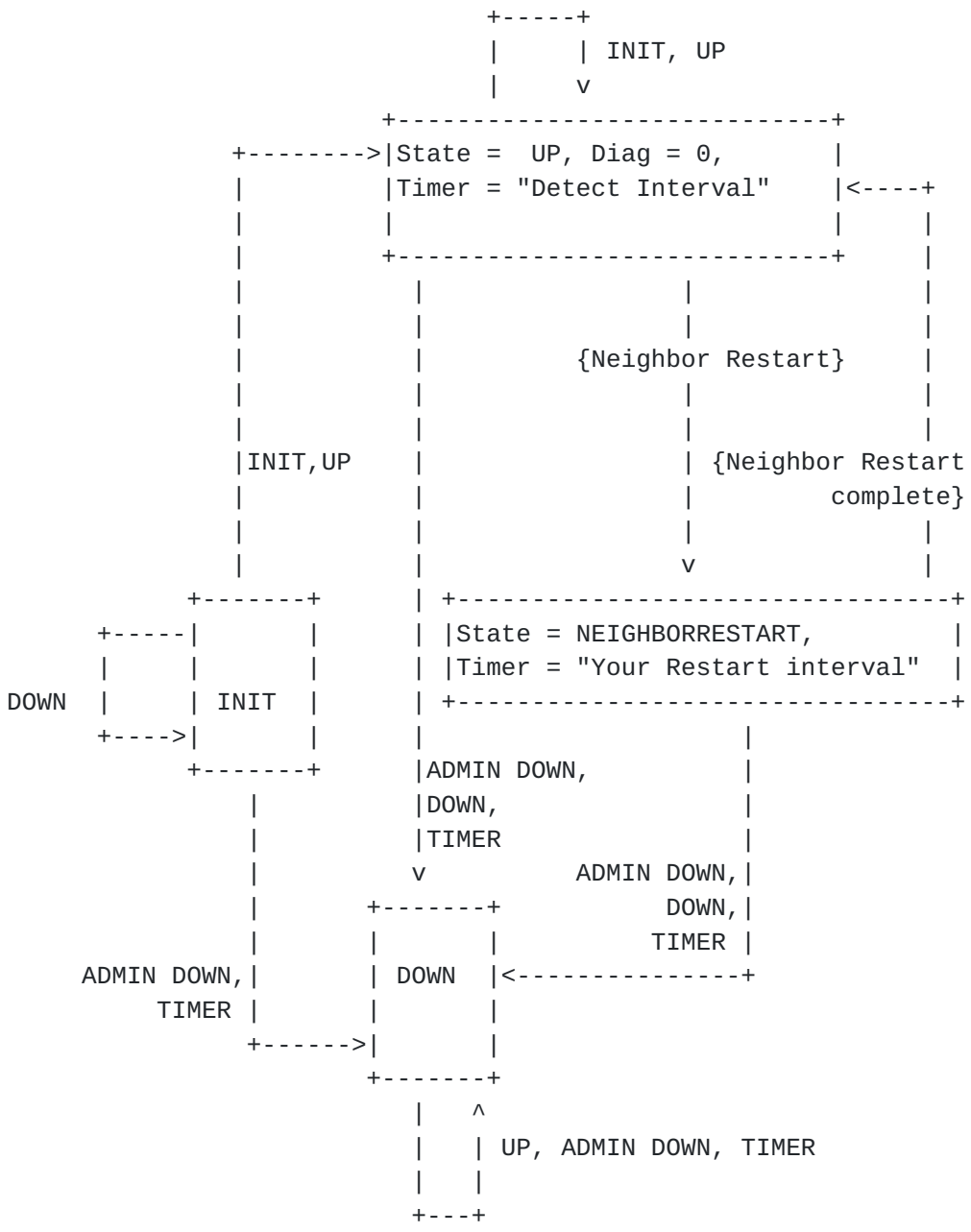
Each system communicates its session state in the State (Sta) field of the BFD Control packet, and that received state, in combination with the local session state, drives the state machine.Please refer to [BFD] for state the machine diagram and detailed explanations of the state transitions.

The following diagram provides an overview of the state machine, for state transitions for BFD with GR support (where "Your Restart Interval" has a non-zero value greater than Detection time). This document introduces a new state, "NeighborRestart" to the BFD state machine. The use of that new state has not been adopted by the BFD working group and should not be included in implementations.

Furthermore, as noted in Section 4.2, no mechanism exists or has been proposed for communicating this additional state to a BFD peer. Therefore, it must be recognized that the state machine shown here is purely hypothetical and that all procedures described do not form part of the BFD specification.

The "Your Restart Interval" must have a value greater than the detection time value. If this value is zero or less than the detection time value, the state transitions should completely follow the BFD state machine as defined by [BFD].

The notation on each arc represents the state of the remote system (as received in the State field in the BFD Control packet) or indicates the expiration of the Detection Timer.



Note1: This state diagram holds only for BFD with GR extension as described in this document, which implies that "your Restart Interval" has a value greater than the Detection time value of the established session. This state machine must not be implemented.

Note2: The labels of the diagram in braces {} indicate the GR Specific events on the remote neighbor (Restart/Restart complete).

The State Transitions involving the new state, NeighborRestart is explained in the next section of this document.

6. Theory of Operation

This section describes the possible operation of the protocol elements and state machine described in the previous sections. The processes described here were discussed and rejected by the BFD working group. In particular, the processes presented are specific to the GR and are not complete. The BFD Working group does not recommend these processes for implementation and must not form part of a deployed BFD system.

6.1. Session Establishment and GR Timer exchange

The BFD session establishment follows the procedures as described in[BFD]. If the technology described by this document were to be implemented, the BFD control packets would have the following fields with the values given below.

A new section to the BFD control packet format, "My Restart Interval" (Section 4.3) needs to have a non-zero value that is greater than the detection time.

A new section to the BFD control packet format, "Your Restart Interval" (Section 4.4) needs to have a non-zero value that is greater than the detection time.

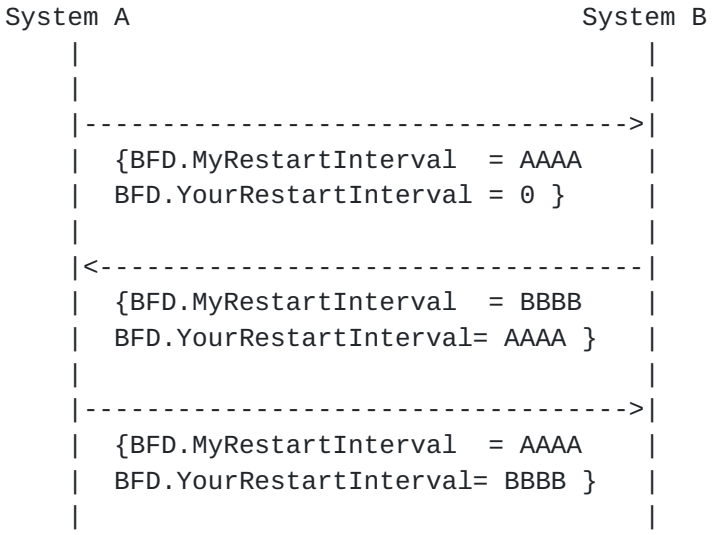
The "My Restart Interval" and "Your Restart Interval" are used in exchanging the GR timers information between the systems.

"My Restart Interval" is the time interval in microseconds, that this system expects its remote system to wait for before bringing down its BFD session with that system.

"Your Restart Interval" is the time interval in microseconds, specified by the remote system, that it expects this system to wait for, before bringing down its BFD session with the remote system.

Once the packet exchanges are complete and the BFD sessions are up,every BFD session will have information about the time interval its remote system will wait during a Restart, and also the time interval this system has to wait when the remote system restarts. The "My Restart Interval" and the "Your Restart Interval" values can be modified after the session is up, just like the other BFD parameters, and in this case the packet exchanges will sync up the restart interval times (My and Your) on both the sides appropriately.

The exchange of GR Specific parameters during BFD session establishment is indicated in the diagram below. The diagram shows only part of control packets, for the purpose of clarity.



The initial BFD packet exchange between local system and remote system will have the exchanged values for the "My Restart Interval" or 0. The "Your Restart Interval" will reflect the value received in My Restart Interval" from the corresponding remote system, or be Zero if that value is not set (value of 0).

A value of Zero for "Your Restart Interval" means that the BFD GR is disabled at the remote end, and similarly a value of Zero for "My Restart Interval" means that BFD GR is disabled at the transmitting system.

6.2. Remote Neighbor Restart and Recovery

When the BFD neighbors have established their BFD sessions (with their BFD GR timer values exchanged as described above), the following set of operations take place when the remote neighbor attempts a graceful restart (for example, with a GR enabled routing protocol like OSPFv2/IS-IS tied with BFD).

Once the packet exchanges are complete and the BFD sessions are up, every BFD session will have information about the time interval its remote system will wait during a Restart, and also the time interval this system has to wait when the remote system restarts.

For clarity, let us revisit the BFD timers and BFD detection time as described in [BFD].

The Detection Time (the period of time without receiving BFD packets after which the session is determined to have failed) is not carried explicitly in the protocol. Rather, it is calculated independently in each direction by the receiving system based on the negotiated transmit interval and the detection multiplier.

This means that a BFD control packet should be received from the remote neighbor within the detection time. When the BFD control packet is not received from the remote neighbor within this time, the timer expiry should bring the BFD session state to down.

During a Graceful, we may end up in a situation that the routing protocol (like OSPFv2) is in graceful restart mode with the remote neighbor restarting, and the system not receiving BFD control packets within the detection time, due to other CPU intensive processes in the system. The technology described by this document addresses this issue.

If the set of systems had their BFD sessions established, with GR support as described in this document, when the remote neighbor restarts it will set the BFD diagnostics field to a value to indicate "Neighbor Restarting" in the control packet to its neighbor (local system).

When the local system receives a BFD control packet with its diag field set to indicate "Neighbor Restarting", the local system will update its timer to the previously exchanged value of "Your Restart Interval".

This effectively means that the local system should wait for a BFD control packet for "Your Restart Interval" instead of Detection time. This will be the case as long as the diag field from the remote neighbor indicates Neighbor Restart. The BFD moves from "Up" to "NeighborRestart" state.

7 Security Considerations

The security implications of the ideas discussed in this document have not been examined. It is likely that the security considerations discussed in [BFD], [BFD-1HOP] apply to this document.

8 IANA Considerations

This document is a historical record of the work and the discussions on the BFD working group, on a possible solution to Graceful Restart. No IANA action is requested.

9 Acknowledgments

The Author likes to acknowledge caldwel.E and Ramesh.M for their early inputs to this document. Thanks to Fred Baker, Dave Ward (bfd-wg Chair), Senthil Sivakumar, Nevil Brownlee (RFC-ISE) and Adrian Farrel (routing AD) for valuable advice, help and guidance.

The Author also likes to thank Joel Bion, Karthik, Mridhula, Ramya, and members of ISOC fellowship committee for their wholehearted

support towards the author's IETF activities.

10 References

10.1 Normative References

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

[BFD-1HOP] Katz, D., and Ward, D., "BFD for IPv4 and IPv6 (Single Hop)", RFC 5881, June, 2010.

[RFC5882] Katz, D. and Ward, D., "Bidirectional Forwarding Detection", RFC 5882, June 2010

10.2 Informative References

[IS-IS-GRACE] Shand, M., and Ginsberg, L., "Restart signaling for IS-IS", RFC 5306, October 2008.

[OSPF-GRACE] Moy, J., et al, "Graceful OSPF Restart", RFC 3623, November 2003.

[RFC5226] T. Narten, H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC5226, May 2008.

Authors' Addresses

Palanivelan Appanasamy
Principal Software Engineer,
Networking/WAN, EMC Corporation,
Bangalore-560048.
India.

EMail: Palanivelan.Appanasamy@emc.com