Network Working Group
Internet Draft
Intended status: Proposed Standard
Expires: May 2011

S. Giacalone
Thomson Reuters
November 15, 2010

A. Soliman
Thomson Reuters
November 15, 2010

Bidirectional Forwarding Detection (BFD) Express Path
draft-giacalone-bfd-express-path-00.txt

Abstract

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance criteria (e.g. latency) have become (or are becoming) as (or more) critical to data path selection than other metrics.

This document describes extensions to the BFD protocol, such that network performance information can be gathered in a scalable fashion, and subsequently used (by other protocols) to make path selection decisions. These extensions will also provide granular performance monitoring information.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on May 15, 2011.

Table of Contents

1. Introduction

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance information (e.g. latency) have (or are becoming) as (or more) critical to data path selection than other metrics. In many of these networks, bandwidth is relatively rich and homogeneous (e.g. a WAN core network of all 10 or 20 Gigabit Ethernet links, or greater), however path length (and therefore latency) can vary in between nodes, and can even change based on the path protection scheme used. In these networks, extremely large amounts of money rest on the ability to make trades faster than the competition, and the ability to access real time market data. In certain financial services networks, hop count, cost, and bandwidth are only tangentially important. Rather, it would be beneficial to be able to granularly monitor network performance and/or make path selection decisions based on performance data (such as latency) in a cost-effective and scalable way.

Many performance sensitive networks have already implemented BFD [RFC5880] to improve convergence time. This document describes extensions to the BFD protocol (hereafter called “BFD Express Path”), that can be used to derive latency (and other performance information). BFD Express Path can provide performance data for every link or path where BFD is enabled. BFD Express Path is open, scalable, and provides a number of benefits:

- Once BFD Express Path gathers latency information, it could subsequently be distributed using extensions to existing routing protocols, such as OSPF [RFC2328] and OSPF-TE [RFC3630]. In this scenario, BFD Express Path would be used with BFD on a link-by-link basis [RFC5881]. Using this link-by-link data, end-to-end path selection can be performed based on latency metrics, as part of the normal operation of various routing protocols (e.g. by replacing cost with latency) or by using “second order” control plane protocols such as CSPF, RSVP-TE [RFC3209], etc.

Note that routing protocol extensions are out of the scope of this document, and will be covered elsewhere.

Although this document focuses on latency, there is no reason why implementations of BFD Express Path cannot be used to gather jitter and loss information. Note, however, that with respect to path selection protocol interactions and support, this document only focuses on latency.

- Since BFD Express Path operates between pairs of nodes, it can create a scalable, open mechanism for link-by-link SLA compliance monitoring, which is an important issue in large, diverse networks that use transport services from various providers. In networks like this, end-to-end latency is not always useful for SLA enforcement (since various links from different providers may make up a path). This link-by-link performance monitoring data could easily be gathered by looking at a routing protocol's state database (on any router in an area, depending on what is being monitoring and disseminated by the routing protocol), using SNMP [RFC1441] on a per device basis, or in other ways.

- In addition to looking at link-by-link latency, BFD Express Path can also be used to understand overall path latency (or other parameters). To do this, BFD Multihop [RFC5883] could be used to take measurements directly, or more simply, the topology database from any device participating in an extended routing protocol that distributes BFD Express Path information could be consulted. Using this information, it would be possible to bring links or MPLS TE tunnels out of service (if needed, based perhaps on SLA), reroute traffic, or take other actions. This is particularly useful in networks or scenarios where an end-to-end service that breaches an SLA (like a latency SLA) is considered "down", even though the network forwarding and control planes are both "up".

- In addition to enhanced routing and SLA management, BFD Express Path's link-by-link or end-to-end network performance information can be used to enable threshold based alerting (network management) in a cost-effective and scalable fashion. In large, diverse financial networks, it is critical to know when performance (e.g. latency) between nodes, points, or services varies.

- BFD Express Path is simple to deploy, easy to monitor, and scalable to provision. It is low cost, integrated with existing protocols, and does not require expensive tools, hardware, or

systems. BFD Express Path provides link-local and end-to-end functionality and is an open protocol for performance and latency based monitoring, analysis, and routing.

2. Conventions used in this document

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. Express Path Extensions to BFD

The main change BFD Express Path makes to BFD [RFC5880] is the addition of timestamp fields. The basic structure and intent of these timestamps are similar to ICMP Timestamp [RFC792]. The format of these fields, relative to the basic BFD control packet will be explained in sections 4 and 5.

To support BFD Express path, it is also proposed that 2 new diagnostic codes [RFC5880] be assigned. The first diagnostic code indicates that the sender is using BFD Express Path. The second indicates whether the sender is synchronizing clock to a clock synchronization protocol (e.g. NTP [RFC1305]) source. The first code (hereafter called the “Express Path” bit) is required to delimit and differentiation the packet extensions. The second code (hereafter called the “clock sync” bit) ensures that similar clock accuracy is used between systems.

Note that because of the limitations of NTP with respect to clock accuracy, this document does not preclude the use of other time synchronization protocols, which may provide more accurate synchronization. The use of other clock synchronization protocols may have implications on the diagnostics codes assigned, however.

4. Originating BFD Express Path Packets

When originating a BFD Express Path Control packet, the sender MUST append the "Originate Timestamp" field to the normal BFD Control packet [RFC5880], as shown below. In the diagram below, the first 24 bytes of the BFD packet are left unchanged from [RFC5880], except for the assignment of new diagnostic codes, adjustment of the packet "length" field to account for the appended field, and the new timestamp field itself:

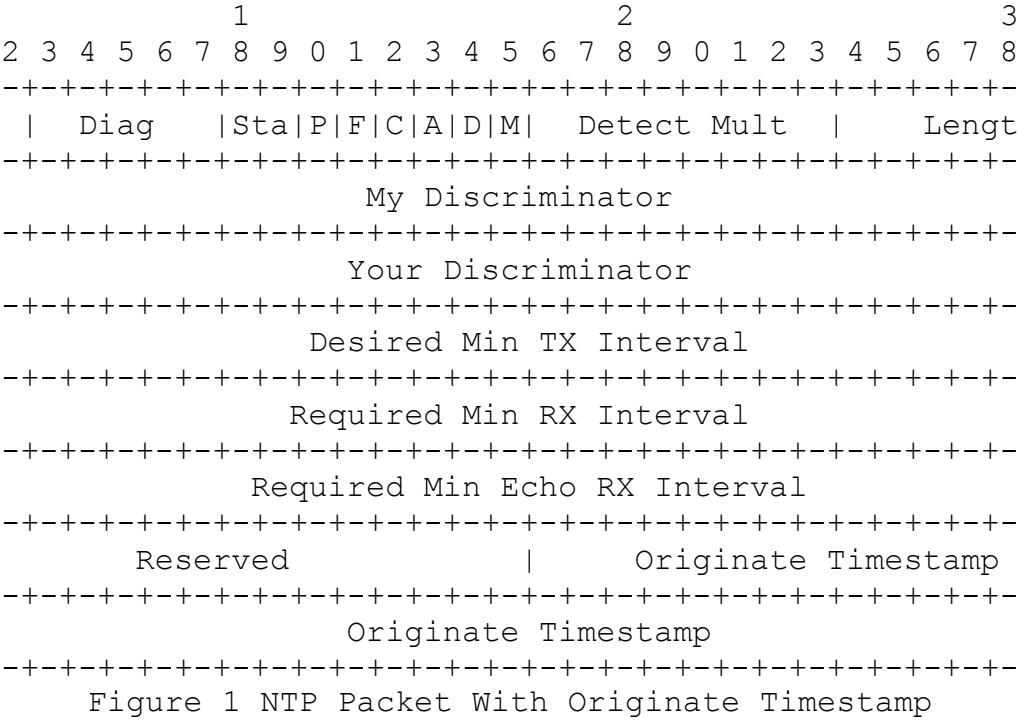

Figure 1 NTP Packet With Originate Timestamp

4.1. Diagnostic Codes

As detailed in section 3, it is proposed that 2 new BFD diagnostic codes be assigned for BFD Express Path. The first diagnostic code indicates that the sender is using BFD Express Path. The second indicates whether the sender is synchronizing clock to a clock synchronization source. The first code is required to delimit and

differentiation and the packet extensions. The second code ensures that similar clock accuracy is used between systems.

4.2. Length

This field specifies the length of the BFD Control packet, in bytes as per [RFC5880]. In this case, the length would be 32.

4.3. Reserved

A 16 bit field, reserved for future use

4.4. Originate Timestamp Field

The Originate Timestamp field is 48 bits in length. The exact format and range of values in the field is TBD. Whatever format is decided, it is recommended that it permit accuracy to the microsecond (uS).

As in ICMP Timestamp [RFC792], the time value in this field MUST represent the time the sender last touched the message before sending it.

Whether or not this timestamp is fixed to some reference time value, such a midnight, as in ICMP Timestamp [RFC792], or relative to a local clock (likely using NTP) is TBD. Space has been allocated to support a reference value in ms (as per [RFC792], plus the uS increment since the last ms “tick”. If more efficient encodings are agreed, this field may be shortened in future draft revisions.

5. BFD Express Path Response Packets

When responding to a BFD Express Path Control packet, the responding system MUST append the “Receive Timestamp” and the “Transmit Timestamp” fields to the BFD Express Path packet, unless BFD Asynchronous Mode [RFC5880], is being used and the “Poll” (P) Bit [RFC5880] is cleared. By removing the requirement for additional timestamps in pure Asynchronous Mode, packet length is reduced, and protocol efficiency is increased.

The diagram below shows a BFD Express Path response packet for a session that is using Echo Mode or where the P bit is set, and therefore, the Receive Timestamp and the Transmit Timestamp and included:

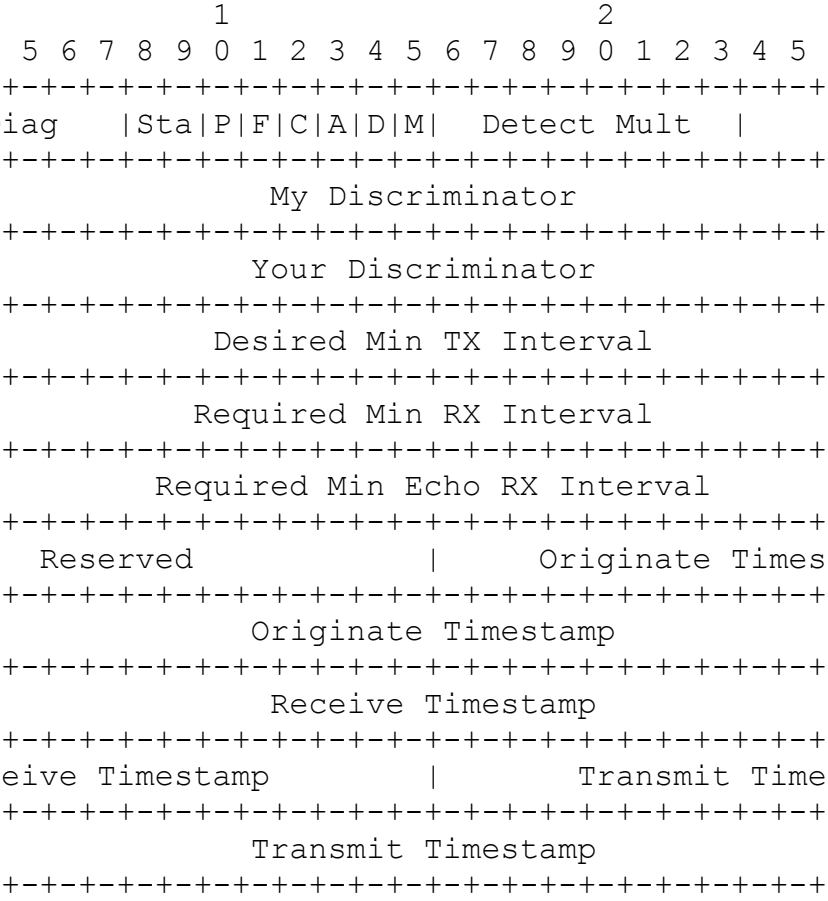

Figure 2 BFD Express Path Response Packet

5.1. Diagnostic Codes

The assignment of diagnostic codes for BFD Express Path is outlined in 3. Note that the responding system MUST set the diagnostic codes according to the setting of the local system. In other words, if the originator has the NTP bit set, but the responder does have NTP enabled, the responder MUST clear the NTP bit.

5.2. Length

This field specifies the length of the BFD Control packet, in bytes [RFC5880]. In this case, the length would be 44.

5.3. Reserved

A 16 bit field, reserved for future use

5.4. Receive Timestamp Field

The receive timestamp field is a 48 bits in length. The exact format and range of values in the field is TBD. Whatever format is decided, it is recommended that it permit accuracy to the microsecond (uS).

As in ICMP Timestamp [RFC792], the Receive Timestamp is the time the responding system first touched the BFD Express Path Packet on receipt.

Whether or not this timestamp is fixed to some reference time value, such a midnight, as in ICMP Timestamp [RFC792], or relative to a local clock (likely using NTP) is TBD. Space has been allocated to support a reference value in ms (as per [RFC792], plus the uS increment since the last ms "tick". If more efficient encodings are agreed, this field may be shortened in future draft revisions.

5.5. Transmit Timestamp Field

The transmit timestamp field is a 48 bits in length. These exact format and range of values in the field is TBD. Whatever format is decided, it is recommended that it permit accuracy to the microsecond (uS).

As in ICMP Timestamp [RFC792], the Transmit Timestamp is the time the responding system last touched the message on sending it.

Whether or not this timestamp is fixed to some reference time value, such a midnight, as in ICMP Timestamp [RFC792], or relative to a local clock (likely using NTP) is TBD. Space has been allocated to

support a reference value in ms (as per [RFC792], plus the uS increment since the last ms "tick". If more efficient encodings are agreed, this field may be shortened in future draft revisions.

6. BFD Mode Support

All implementations of BFD Express Path MUST support Asynchronous mode BFD with NTP (at a minimum) as the clock source.

When other BFD modes (e.g. Echo, Demand, Poll Request, etc) are supported by the BFD implementation, it is RECOMMENDED that BFD Express Path be supported for the other BFD modes as well.

Implementations that support Multihop BFD [RFC5883] SHOULD also support BFD Express Path, Multihop BFD.

7. Error Detection

If the local system has the BFD Express Path bit set in the diag code field, and the remote system does not, an error MUST be generated, as this could lead to an inconsistent topology database, if and when latency information is used for path selection.

If the local system has the Clock Sync bit set in the diag code field, and the remote system does not, an error SHOULD be generated.

Whether or not either of these errors affect the state of the BFD session (i.e. bring it down) is OPTIONAL and implementation specific, however is it RECOMMENDED that this be configurable.

8. Latency, Jitter, and Loss

Implementations of BFD Express Path MUST support latency monitoring. Implementations MAY also support jitter and loss monitoring. The exact algorithms for performing this monitoring and analysis are implementation specific.

9. Sampling and Monitoring

The exact frequency with which BFD Express Path measurements are taken, the sampling mechanisms used, the averaging algorithms used, and the amount of trending data stored are implementation specific, and are out of the scope of this document.

Implementations are encouraged to use “common sense” averaging frequencies for monitoring (e.g. a 30 second average, 1 minute average, 5 minutes, 10 minutes, etc). However, note that from a practitioner’s perspective, the capability to sample at more granular intervals and the ability to store more trending data are generally considered good things, and implementers are encouraged to consider advantages in this area a competitive differentiator.

It also is RECOMMENDED that implementations permit latency measurements to be taken at any particular time (i.e. a snapshot in time).

No matter the sampling and averaging frequencies used, implementations MUST permit configuration of sampling or averaging timers such that users can configure how subsequent path selection algorithms will respond to changes in the performance of the network.

10. Dissemination of Latency Information

The need to gather BFD Express Path data is critical. It would make sense to expose a MIB Variable(s) to do this; however, MIB extensions are currently outside the scope of this document.

The injection of BFD Express Path information into routing protocols is an important application of BFD Express Path, and will be addressed outside the scope of this document.

11. Security Considerations

Security considerations discussed in [BFD], [BFD-1HOP] apply to this document.

12. IANA Considerations

BFD Express Path will require 2 Diagnostic Codes to be assigned as per section 3. In addition, 3 new fields will be added to the BFD [RFC5880] packet; the Originate Timestamp, the Receive Timestamp, and the Transmit Timestamp.

13. References

13.1. Normative References

[RFC792] Postel, J., "INTERNET CONTROL MESSAGE PROTOCOL", RFC 792, September 1981.

[RFC1305] Mills, D. L., Network Time Protocol (Version 3) Specification, Implementation and Analysis, RFC 1305, March 1992

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

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

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

13.2. Informative References

[RFC2328] Moy, J, "OSPF Version 2”, RFC 2328, April 1998

[RFC1441] Case, J., McCloghrie, K., Rose, M., Waldbusser, S., "Introduction to version 2 of the Internet-standard Network Management Framework", April 1993

[RFC2370] Coltun, R., "The OSPF Opaque LSA Option", RFC 2370, July 1998.

[RFC3031 Rosen, E., Viswanathan, A., Callon, R., “Multiprotocol Label Switching Architecture”, January 2001

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3630] Katz, D., Kompella, K., Yeung, D., "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

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

14. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Authors’ Addresses

Spencer Giacalone
Thomson Reuters
195 Broadway, New York NY 10007

Phone: (646) 822 3000
Email: Spencer.giacalone@thomsonreuters.com

Ayman Soliman
Thomson Reuters
195 Broadway, New York NY 10007

Phone: (646) 822 3000
Email: ayman.soliman@thomsonreuters.com