

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 17, 2014

T. Tsou
Huawei Technologies (USA)
B. Li
C. Zhou
Huawei Technologies
J. Schoenwaelder
Jacobs University Bremen
R. Penno
Cisco Systems, Inc.
M. Boucadair
France Telecom
February 13, 2014

**DS-Lite Failure Detection and Failover
draft-tsou-softwire-bfd-ds-lite-06**

Abstract

In DS-Lite, the tunnel is stateless, not associated with any state information, and the CGN function at the AFTR is stateful. Currently, there is no failure detection and failover mechanism for both stateless tunnel and stateful CGN function, which makes it difficult to manage and diagnose if there is a problem. This draft analyzes the applicability of some of the possible solutions.

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 August 17, 2014.

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 [3](#)
- [2.](#) Terminology [3](#)
- [3.](#) Failover Mechanisms [3](#)
 - [3.1.](#) Anycast Approach [4](#)
 - [3.2.](#) VRRP Approach [4](#)
- [4.](#) Solutions [4](#)
 - [4.1.](#) Bidirectional Forwarding Detection (BFD) [4](#)
 - [4.1.1.](#) DS-Lite Scenario [5](#)
 - [4.1.2.](#) Parameters for BFD [5](#)
 - [4.1.3.](#) Elements of Procedure [6](#)
 - [4.1.4.](#) BFD for NAT failure detection [6](#)
 - [4.1.5.](#) Implementation Considerations [6](#)
 - [4.2.](#) Port Control Protocol (PCP) [7](#)
 - [4.3.](#) ICMP Echo Request / Echo Reply (PING) [7](#)
 - [4.4.](#) Comparison of Different Solutions [8](#)
- [5.](#) State Synchronization and Session Re-establishment [8](#)
- [6.](#) IANA Considerations [9](#)
- [7.](#) Security Considerations [9](#)
- [8.](#) Acknowledgements [9](#)
- [9.](#) References [9](#)
 - [9.1.](#) Normative References [9](#)
 - [9.2.](#) Informative References [10](#)
- Authors' Addresses [10](#)

1. Introduction

In DS-Lite [[RFC6333](#)], the IPv4-in-IPv6 DS-Lite tunnel is stateless, no status information about the tunnel is available, and no keep-alive mechanism is available. It is difficult to know whether the tunnel is up or down; and if there is a link problem, the Basic Bridging BroadBand (B4) element can not automatically switch to another Address Family Transition Router (AFTR) so as to continue the network service automatically, without the involvement of operators. Besides, In DS-Lite [[RFC6333](#)], the CGN function at the AFTR is stateful and there is no mechanism to detect whether the NAT44 CGN is functioning in the AFTR. These will create problems for network operation and maintenance.

Possible solutions for failure detection include the usage of Bidirectional Forwarding Detection (BFD), the Port Control Protocol (PCP), and ICMP Echo Request / Echo Reply (PING). The properties of these solutions are discussed in this document and guidelines are provided how to implement failure detection and automatic failover.

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. Terminology

AFTR: Address Family Transition Router.

B4: Basic Bridging BroadBand.

BBF: BroadBand Forum.

BFD: Bidirectional Forwarding Detection.

CPE: Customer Premise Equipment (i.e., the DS-Lite B4).

FQDN Fully Qualified Domain Name.

PCP Port Control Protocol.

3. Failover Mechanisms

The FQDN of the AFTR is sent to the B4 element via a DHCP option, as defined in [[RFC6334](#)]. Multiple IP addresses can be configured for the FQDN of an AFTR on the DNS server. If a B4 element detects a failure on the link to the AFTR, the B4 element MUST terminate the

current DS-Lite tunnel, choose another AFTR address in the list, and create a tunnel to the new AFTR. If necessary, the B4 element SHOULD re-configure the connectivity test tool accordingly and restart the test procedures.

3.1. Anycast Approach

Anycasts may also be used for failover. But there is an ICMP-error-message problem with anycast, that is, when a packet is sent from the AFTR to a B4 element, if one of the routers along the path generates an ICMP error message, e.g., Packet Too Big (PTB), then the error message may not be sent back to the source AFTR but to another AFTR.

There's also a problem with anycast for stateful CGN/AFTR. If there is an asymmetric path though the CGNs, then return path traffic will be dropped as there is no corresponding state table entry in the AFTR.

3.2. VRRP Approach

For active/passive HA in NAT gateways, it's quite common to have a single virtual address offered by VRRP (or a proprietary equivalent) that the upstream routers will use as their next hop. In the event that the master CGN fails, the standby takes over the virtual L3 address. If a VRRP based virtual address is used as the tunnel endpoint, then the clients wouldn't need to be aware of the failover.

4. Solutions

4.1. Bidirectional Forwarding Detection (BFD)

Bidirectional Forwarding Detection [[RFC5880](#)] (BFD) is a mechanism intended to detect faults in a bidirectional path. It is usually used in conjunction with applications like OSPF, IS-IS, for fast fault recovery and fast re-route [[RFC5882](#)]. BFD is being made mandatory for keep-alive for subscriber sessions, including DS-Lite, by the BroadBand Forum (BBF) [[WT-146](#)].

BFD can be used in DS-Lite, by creating a BFD session between the B4 element and the AFTR to provide tunnel status information. If a fault is detected, the B4 element can try to create a DS-Lite tunnel with another AFTR and terminate the existing one, so as to continue network service. BFD could also be used to detect the CGN state at the AFTR, but the detection should be based on per-user.

[I-D.vinokour-bfd-dhcp] proposes using a DHCP option to distribute BFD parameters to B4 elements. But in case of DS-Lite, some of the

key BFD parameters are already available (e.g., peer IP address), and other parameters can be negotiated by BFD signaling or statically configured, so that no extra DHCP option(s) need to be defined.

4.1.1. DS-Lite Scenario

In DS-Lite [RFC6333], the BFD packet SHOULD be sent through an IPv4-in-IPv6 tunnel, as shown in Figure 1. The IPv4 addresses of the B4 element and the AFTR SHOULD be the endpoints of a BFD session.

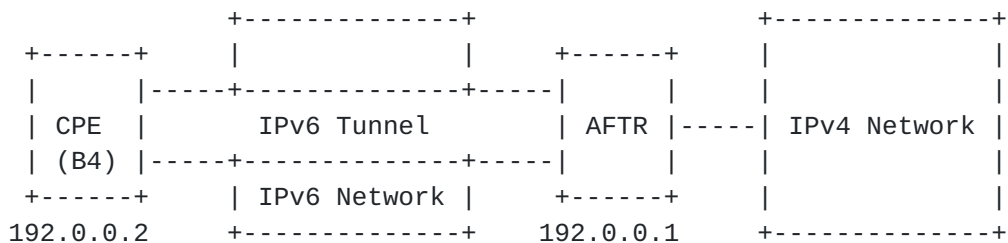


Figure 1: DS-Lite Scenario

4.1.2. Parameters for BFD

In order to set up a BFD session, the following parameters are needed, as shown in Section 4.1 of [RFC5880]:

- o Peer IP address
- o My Discriminator
- o Your Discriminator
- o Desired Min TX Interval
- o Required Min RX Interval
- o Required Min Echo RX Interval

B4's WAN-side IPv4 address is the well-known address 192.0.0.2, and the AFTR's well-known IPv4 address is 192.0.0.1, as defined in section 5.7 of [RFC6333]. The B4 element needs to create an IPv6 tunnel to an AFTR so as to get network connectivity to the AFTR, and send IPv4 BFD packets through the tunnel to manage it.

The other parameters listed above can be negotiated by BFD signaling, and initial values can be configured on B4 elements and AFTRs.

4.1.3. Elements of Procedure

When a B4 element gets online, it will be assigned an IPv6 prefix or address, and also the FQDN of the AFTR, as defined in [RFC6334]. The B4 element will create an IPv6 tunnel to the AFTR with which the B4 element can initiate a BFD session to the AFTR. BFD packets will be sent through the DS-Lite tunnel. As defined in [section 4 of \[RFC5881\]](#), BFD control packets MUST be sent in UDP packets with destination port 3784, and BFD echo packets MUST be sent in UDP packets with destination port 3785.

When sending out the first BFD packet, the B4 element can generate a unique local discriminator, and set the remote discriminator to zero. When the AFTR receives the first BFD packet from a B4 element, the AFTR will also generate a corresponding local discriminator, and put it in the response packet to the B4 element. This will finish the discriminator negotiation in the B4 to AFTR direction, without any manual configuration.

When an AFTR receives the first packet from a B4 element, the AFTR will get the IPv6 address and discriminator of the B4 element, so that the AFTR can initiate the BFD session in the other direction and a similar discriminator negotiation can be carried out.

4.1.4. BFD for NAT failure detection

B4 creates PCP mapping. BFD at AFTR uses an external public interface (or another external mapping) to send a BFD packet to the public PCP mapping created by B4. In this case, the AFTR BFD packet will have a public source IP of interface, which will go through the NAT, therefore exercising the NAT function. B4 will reply to the AFTR external interface.

4.1.5. Implementation Considerations

BFD is usually used for quick fault detection, at a very small time scale, e.g. milliseconds. But in DS-Lite, it may not be necessary to detect faults in such a short time. On the other hand, an AFTR may need to support tens of thousands of B4 elements, which means an AFTR will need to support the same number of BFD sessions. In order to meet performance requirements on an AFTR, it may be necessary to extend the time period between BFD packet transmissions to a longer time, e.g., 10s or 30s.

Compared to other solutions, BFD has a simple and fixed packet format, which is easy to implement by logic devices (e.g., ASIC, FPGA). Complicated protocols are usually processed by software which is relatively slow. An AFTR may need to support 10000-20000 users,

and if the protocol is handled by software, it will bring extra load to the AFTR.

4.2. Port Control Protocol (PCP)

[[RFC6887](#)]PCP is a NAT traversal tool. It can also be used for network connectivity test if PCP is supported in the network. A common use case of PCP is to create a pinhole so that external users can visit the servers located behind a NAT. The lifetime of the pinhole mapping is usually long, e.g., hours, and the lifetime will be refreshed periodically by the client before it is expired. For the purpose of network connectivity tests, a B4 element can create a mapping in the CGN via PCP, with a short life time, e.g., 10s of seconds, and keep on refreshing the mapping before it expires. If any refresh requests fail, the B4 element knows that something is wrong with the link or the PCP server or the CGN.

In order to detect the network connectivity of the DS-Lite tunnel, the encapsulation mode MUST be used for PCP: PCP packets are sent through the DS-Lite tunnel.

PCP can detect the failure of more components of the DS-Lite system. Besides failures of the link and the routing, it also covers NAT functions.

4.3. ICMP Echo Request / Echo Reply (PING)

PING is commonly implemented using the Echo Request and Echo Response messages of the Internet Control Message Protocol (ICMP) [[RFC0792](#)] [[RFC4443](#)]. In case of DS-Lite, a B4 element can send Echo Request packets to the AFTR periodically. If the B4 element does not receive Echo Response packets for a certain number (e.g., 3) of Echo Request packets, then the B4 element decides that a fault has been detected.

In order to test the connectivity of DS-Lite tunnel, Echo Request packets MUST be sent using ICMPv4, rather than ICMPv6.

Since ICMP is an integral part of any IP implementation, the usage of PING to detect tunnel failures does not require any special implementation efforts on the B4 elements. However, on AFTRs that process ICMP messages in software rather than in hardware, the usage of PING might lead to scalability issues.

4.4. Comparison of Different Solutions

	Availability	Packet format	Additional functionality ontop of keepalives	Configuration/ provisioning overheads
BFD	Widely used/ network side, less used/ terminal side	Simple fixed	Bidirectional status synchronization	Similar
PCP	Less than BFD/ICMP	Vari- table	No bidirectional detection	
ICMP	Ubiquitous		Network/CGN initiated detection	

Figure 2: Comparison of different solutions

Figure 2 gives a direct comparison among different solutions. Compared to other solutions, BFD has a simple and fixed packet format, which is easy to implement by logic devices (e.g., ASIC, FPGA). Complicated protocols are usually processed by software which is relatively slow. ICMP is widely used than PCP/BFD, while BFD is more widely used in the router and CGN side than in the terminal side. However, from the aspect of failure detection, BFD has explicit capability of bidirectional status synchronization to guarantee the consistency of the failure status of both sides. ICMP could actively initiate status detection from the network side or CGN side, while PCP could not. PCP has no capability of bidirectional detection. Considering the configuration/provisioning overheads, since there is normally TR-069 server at the network management side. So it is similar for each approach.

From the above comparison, BFD is selected as the failure detection approach in this document.

5. State Synchronization and Session Re-establishment

There should be a state sync mechanism between active AFTR and backup AFTR, to synchronize the state of each user between the two AFTRs. This mechanism is to guarantee that the traffic returning to the B4 is from the backup AFTR, if the service is shifted to that AFTR. The BFD link for both active AFTR and backup AFTR should be set up in the

initial state. When the active AFTR is detected in failure, the service will be shifted to the backup AFTR. If the backup AFTR is detected in failure, it will notify the network management server to fix the failure.

In the hot-standby case, the master AFTR and the backup AFTR will synchronize and backup the session. So there is no need to re-establish the TCP session in the event of an AFTR failure. But in the cold-standby case, if there is an active TCP session through the CGN function of an AFTR, and this AFTR fails, then the TCP session will need to be re-established by the client because only the capability is reserved but the session is not backup.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

In the DS-Lite [[RFC6333](#)] application, the B4 element may not be directly connected to the AFTR; there may be other routers between them. In such a deployment, there are potential spoofing problems, as described in [[RFC5883](#)]. Hence cryptographic authentication SHOULD be used with BFD as described in [[RFC5880](#)] if security is concerned.

8. Acknowledgements

The authors would like to thank Ian Farrer for his valuable comments.

9. References

9.1. Normative References

- [RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, [RFC 792](#), September 1981.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", [RFC 4443](#), March 2006.
- [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.
- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", [RFC 5883](#), June 2010.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", [RFC 6333](#), August 2011.
- [RFC6334] Hankins, D. and T. Mrugalski, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Option for Dual-Stack Lite", [RFC 6334](#), August 2011.
- [RFC6887] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", [RFC 6887](#), April 2013.
- [WT-146] Kavanagh, A., Klamm, F., Boucadair, W., and R. Dec, "WT-146 Subscriber Sessions (work in progress)", Apr 2012.

9.2. Informative References

- [I-D.vinokour-bfd-dhcp]
Vinokour, V., "Configuring BFD with DHCP and Other Musings", May 2008.

Authors' Addresses

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Brandon Li
Huawei Technologies
M6, No. 156, Beiqing Road, Haidian District
Beijing 100094
China

Phone:
Email: brandon.lijian@huawei.com

Cathy Zhou
Huawei Technologies
China

Phone:
Email: cathy.zhou@huawei.com

Juergen Schoenwaelder
Jacobs University Bremen
Campus Ring 1
Bremen 28759
Germany

Phone:
Email: j.schoenwaelder@jacobs-university.de

Reinaldo Penno
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, California 95134
USA

Phone:
Email: repenno@cisco.com

Mohamed Boucadair
France Telecom
Rennes, 35000
France

Phone:
Email: mohamed.boucadair@orange.com

