

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 2 September 2022

G. Qian
T. Zhou
Huawei
1 March 2022

IPv6 Minimum Multipath MTU Detection
draft-qian-6man-ipv6-multipath-mtu-detection-00

Abstract

In current multipath load balancing network scenario, all path detection mechanisms have a defect. A typical load balancing route selection mechanism cannot cover all forwarding paths, which will cause missing detection. This document describes how to extend a new path detection mechanism to instruct intermediate devices to send probe packets to all downstream paths. This new mechanism is named load-sharing multipath replication forwarding (LMRF).

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [BCP 14](#) [[RFC2119](#)] [[RFC8174](#)] when, and only when, they appear in all capitals, as shown here.

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 <https://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 2 September 2022.

Internet-Draft

6man Multipath MTU Detection

March 2022

Copyright Notice

Copyright (c) 2022 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 (<https://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 Revised BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Revised BSD License.

Table of Contents

1.	Introduction	2
2.	Terminology	3
3.	Scenario Description	3
3.1.	Example	3
3.2.	Solution	5
4.	Detail solution	5
4.1.	IPv4 solution	5
4.2.	IPv6 solution	5
4.2.1.	Detection Solution	5
4.2.2.	Modifications to existing mechanisms	6
5.	Supplementary description of the protocol	9
6.	Benefits	10
7.	Acknowledgements	10
8.	IANA Considerations	10
9.	Security Considerations	10
10.	Normative References	10
	Authors' Addresses	11

[1.](#) Introduction

In the current multipath load balancing scenario, a path detection mechanism has a defect. A common load balancing route selection solution cannot cover all forwarding paths, which causes missing detection. This document describes how to extend a new probe mechanism to instruct intermediate forwarding devices to send probe packets to all downstream paths.

Typical problem: During path MTU detection, the path MTU of a path cannot be used as the path MTU of all load balancing paths. In this case, the source selects the minimum path MTU of different paths as the path MTU of the entire path to ensure normal forwarding on the intermediate network.

Currently, there are some solutions in the industry, such as the Paris trace solution. By constructing a large number of packets at the source and modifying information such as the transport-layer port number of the packets, the forwarding device on the network can hash the packets to as many forwarding paths as possible during route selection. This solution cannot ensure that all paths are covered. In addition, a large number of packets need to be constructed at the source, which affects network performance and imposes more workload and skill requirements on O&M engineers.

2. Terminology

The following terminology is used in this document.

MTU: Maximum Transmission Unit

Path MTU: path maximum transmission unit

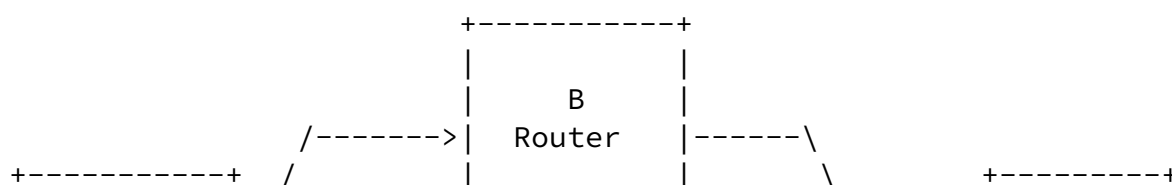
TWAPM: Two-Way Active Measurement Protocol

BFD: Bidirectional Forwarding Detection

LMRF: Load-sharing multipath replication forwarding

3. Scenario Description

3.1. Example



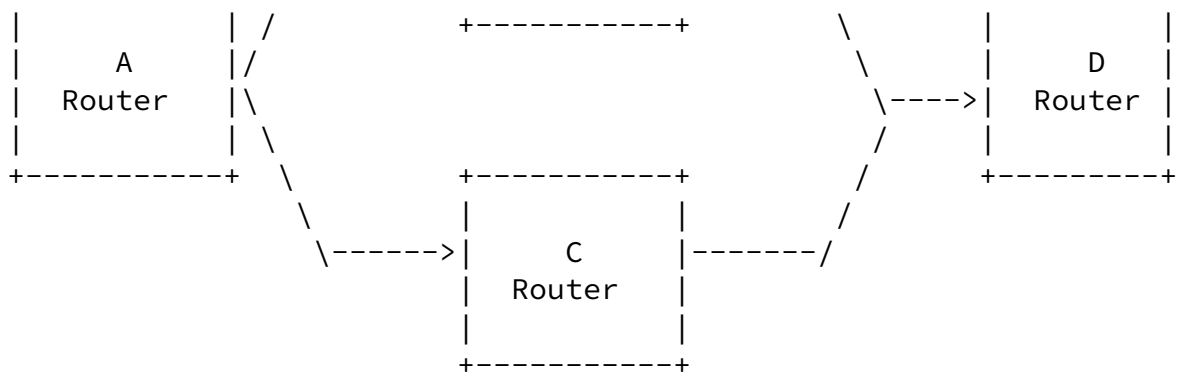
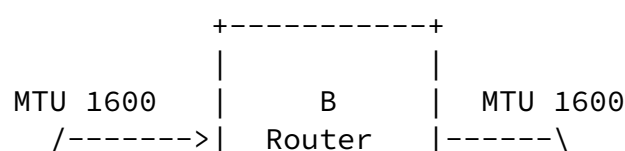


Figure 1: Multihop Network Example

As shown in Figure 1, there are two paths from A to D: A-B-D and A-C-D. The two paths are ECMP paths from A to D. Data packets from A to D are transmitted based on the 5-tuple or triplet information in the packet header. Selects a path based on the hash calculation result. TCP/UDP/ICMP packets are routed based on quintuple, and raw IP packets are routed based on triplet. Take ping packets as an example. The source IP address, destination IP address, protocol number, ICMP type, and ICMP code are used for hash calculation. The result is used for ECMP route selection. Therefore, ping packets from A to D can always cover only one path. Therefore, even if the ping result is normal, services may be abnormal. Conversely, when a service fault occurs, the ping detection may be normal.

Similar problems occur in trace route detection, BFD detection, TWAMP detection, and path MTU detection.

In multi-channel load balancing scenarios, incorrect path MTU detection may cause service exceptions. To simplify packet processing and improve processing efficiency, IPv6 packets are fragmented only on the source node. Therefore, the IPv6 path MTU discovery protocol must be implemented. The latest document ([draft-ietf-6man-mtu-option-11](#) - IPv6 Minimum Path MTU Hop-by-Hop Option) provides the path MTU discovery method for a single path, but does not solve the path MTU problem in multipath scenarios.



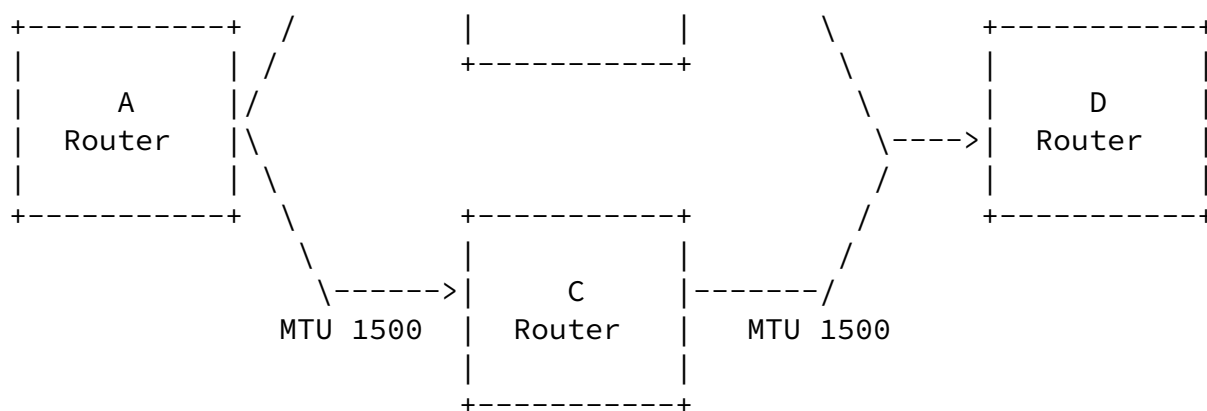


Figure 2:MTU in Multipath Network

As shown in Figure 2, if the path MTU probe packet from A to D is A-B-D, the path MTU of this path is 1600, and the path MTU of the path A-C-D is 1500, Packet loss occurs when data packets with more than 1500 bytes are routed to route A-C-D.

[3.2.](#) Solution

A universal replication detection mechanism is required to support connectivity detection, path MTU detection, and delay detection. This document discusses enhancements to IP header to support multipath detection.

Path MTU detection affects service availability. Therefore, this document focuses on the problem of path MTU detection. Other problems, such as connectivity monitoring and delay monitoring, will be discussed in the future.

[4.](#) Detail solution

[4.1.](#) IPv4 solution

This document focuses on the IPv6 network solution, IPv4 network solution will be discussed in the future.

[4.2.](#) IPv6 solution

[4.2.1.](#) Detection Solution

For IPv6, Hop by hop header and Destination header are extended to carry the multipath replication switch and MTU detection switch. For details, see [section 4.2.2](#). The source node marks the flag, and the intermediate device and tail device perform corresponding processing. After the replication function is enabled on the source node, the source node and transit node copy probe packets to all downstream load balancing paths. After the MTU detection function is enabled on the source node, the source node and intermediate node add the MTU value of the outbound interface to the packet. You can add the MTU value to the packet one by one, or you can compare the MTU value and enter the minimum value. The end node responds to all received detection packets, carries the MTU added along the path, and sends the packets to the source node. The end node can also compare the packets and select the smallest MTU as the final path MTU. To simplify the packet format, packet size, and data-plane processing, it is recommended that only the minimum MTU be reserved in packets. In addition, the path MTU aging mechanism needs to be modified. Considering that the network topology may change, the path MTU may increase. If you always select the minimum value, you can never increase it. Therefore, if no path MTU smaller than or equal to the current path MTU is received for a long time, the current path MTU may be set to an aging state. When the path MTU is in the aging state, the path MTU may be replaced by a larger path MTU.

[4.2.2.](#) Modifications to existing mechanisms

[4.2.2.1.](#) Modification of the packet structure

The hop-by-hop extension header is used in common IP packet. The TTTT needs to be allocated by the IANA.

Option Type	Option Data Len	Option Data
+-----+-----+-----+-----+-----+		
BBCTTTTT 00000011 RRRRRRMD -----MTU-----+		
+-----+-----+-----+-----+-----+		

R:Reserved
M:Path MTU detection flag
D:Load balancing duplicating flag

MTU:Minimum MTU on the path

The reply packet uses the DH extension header, and the TTTT needs to be allocated from the IANA.

Option Type	Option Data Len	Option Data
BBCTTTT	00000010	-----MTU-----

MTU:Minimum MTU on the path

[4.2.2.2.](#) Source node behavior

1. Enable the load balancing duplicating flag.
2. Enable the MTU detection flag.
3. Set the detection timer: The system periodically sends detection packets in duplicate mode and carries the MTU information of its own interface. You are advised to set the timer interval to minutes, which is configurable using the command line.
4. After receiving the response packet from the tail node, the ingress node compares the path MTU value with the local path MTU value and selects the minimum value.
5. Set the path MTU aging timer: The lifetime of the path MTU is periodically updated. When a smaller path MTU or equivalent path MTU is received, the timer is cleared. It is recommended that the timer be set to three times of the detection timer.

6. When the path MTU aging timer expires, the path MTU is set to the aging state and the minimum MTU detected in the next detection period is used to overwrite the path MTU.

[4.2.2.3.](#) transit node behavior

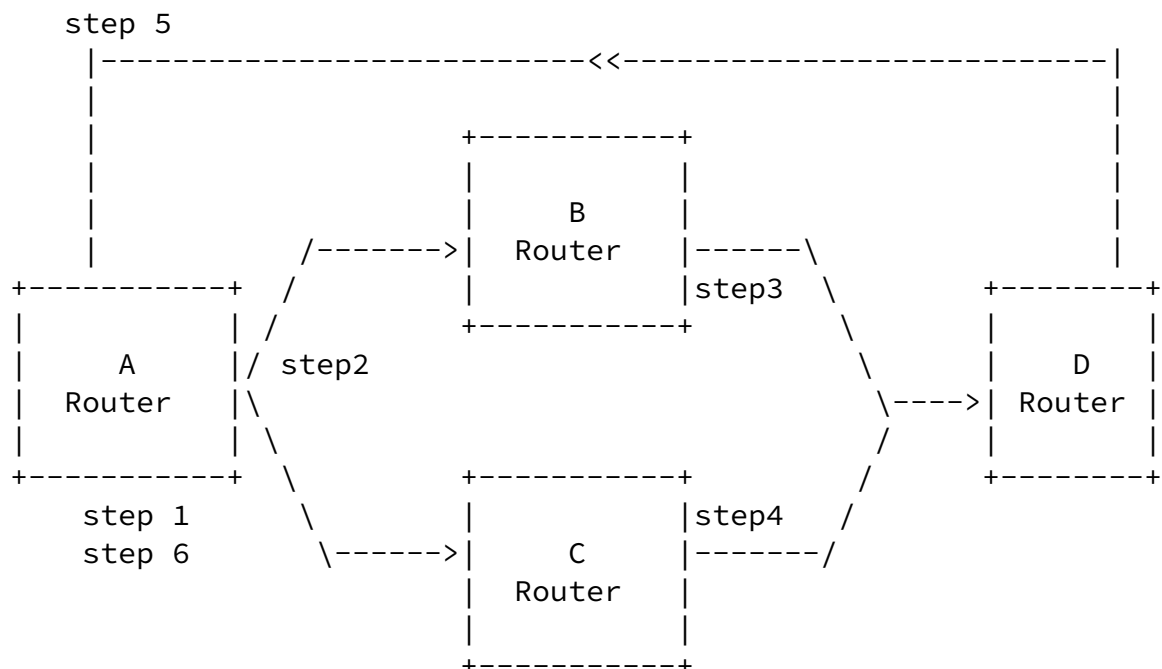
1. Duplicating is performed to all load balancing next hops based on the enabling flag of the load balancing duplicating flag.

2. Compare the MTU in packet with the local output interface MTU, and replace the MTU in the packet with the smaller one.

4.2.2.4. Destination node behavior

1. Send Reply to source node according to all received packets and fill back MTU value get from the received packets.

4.2.2.5. Process flow



step 1. Router A try to dicovery the path mtu to Router D

step 2. Two packets will be send to Router D through Router B and Router C, A-B-D path MTU set as 1600, A-C-D path MTU set as 1700

step 3. Router B received the packet and transfer to Router D, and modify the MTU to 1500

step 4. Router C received the packet and transfer to Router D, and

modify the MTU to 1600

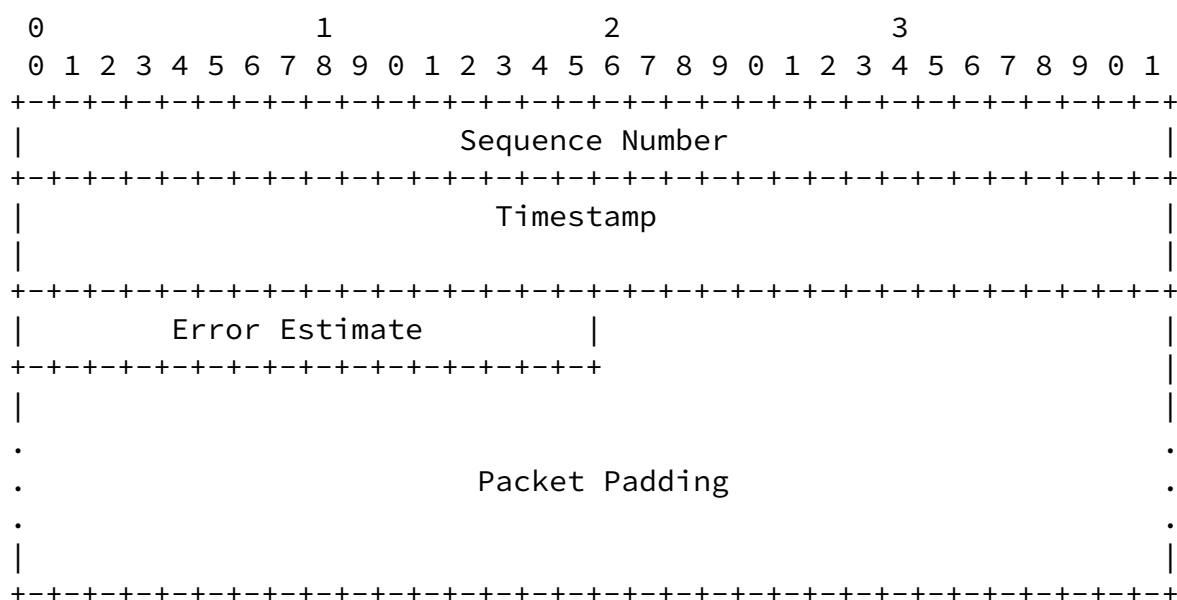
step 5. Router D received two packets and reply to Router A with the corresponding path MTU

step 6. Router A updates local Path MTU with 1500, which is the smallest one among all reply packets.

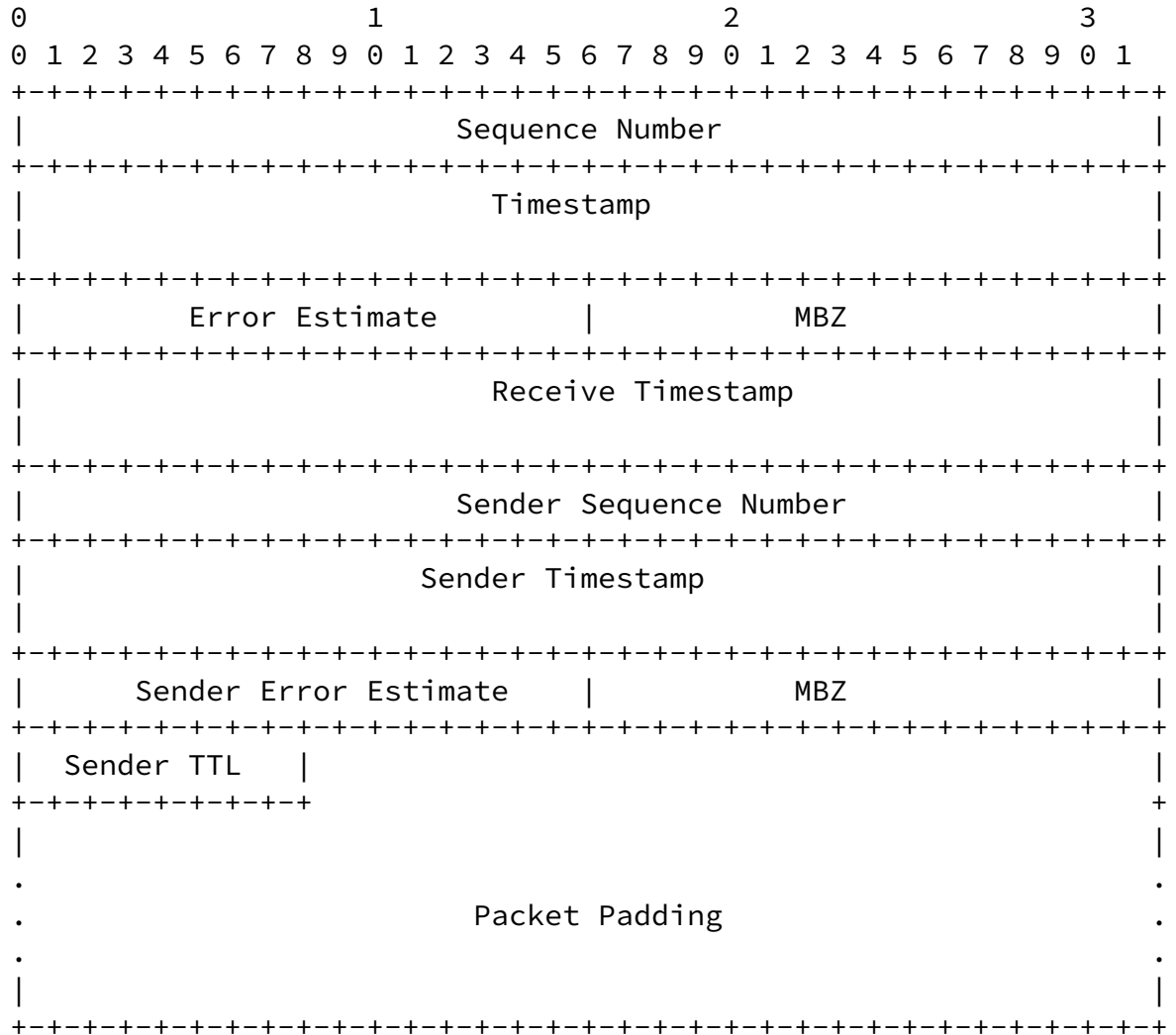
[4.2.2.6.](#) Uplayer protocol consideration

This function does not depend on upper-layer protocols and can work with any upper-layer protocols, such as TCP, UDP, ICMP, Quic, and TWAMP.

Take TWAMP as an example, TWAMP-test packets carry hop-by-hop extension headers and enable M and D flags to detect the MTU of multipath. Sequence numbers are used to identify multiple copies of a packet.



The receiver replies to the source as follows:



Sender Sequence Number is a copy of the Sequence Number of the packet transmitted by the Session-Sender that caused the Session-Reflector to generate and send this test packet.

5. Supplementary description of the protocol

1. In SDN scenarios, path MTUs can be sent to the controller by telemetry, and controller then transfer the packets to source node. This is not discussed in this document.
2. The detection protocol can be extended by TWAMP, BFD, or other OAM protocol. This document does not provide any analysis.
3. This solution assumes all devices on the network support this solution. If intermediate devices do not support, real path MTU will be not detected, Then, PTB will be used to detect the path MTU.

4. The detection of connectivity faults and parameters such as latency in multipath load balancing scenarios will be discussed in future.

6. Benefits

This solution provides accurate path MTU detection in load balancing scenarios to prevent packet loss caused by excessively large packets.

7. Acknowledgements

Thank you to Yang Pingan, Zhao Ranxiao, Xia Yang, Wu Qin, Yudan, and others for participating in the solution discussion and helping improve the solution.

8. IANA Considerations

For carrying the Load balancing duplicating flag and Path MTU detection flag, new option types need to be defined in the existing RH and Hop by Hop headers.

9. Security Considerations

Considering the impact of packet replication on device and network performance, packets in replication mode need to be traced, encrypted, URPF, security filtering, and rate limiting.

10. Normative References

[I-D.ietf-6man-mtu-option]

Hinden, R. M. and G. Fairhurst, "IPv6 Minimum Path MTU Hop-by-Hop Option", Work in Progress, Internet-Draft, [draft-ietf-6man-mtu-option-12](https://www.ietf.org/archive/id/draft-ietf-6man-mtu-option-12), 27 January 2022, <<https://www.ietf.org/archive/id/draft-ietf-6man-mtu-option-12.txt>>.

- [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", [RFC 1191](#), DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/info/rfc1191>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

Qian & Zhou

Expires 2 September 2022

[Page 10]

Internet-Draft

6man Multipath MTU Detection

March 2022

- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", [RFC 2460](#), DOI 10.17487/RFC2460, December 1998, <<https://www.rfc-editor.org/info/rfc2460>>.
- [RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", [RFC 4821](#), DOI 10.17487/RFC4821, March 2007, <<https://www.rfc-editor.org/info/rfc4821>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in [RFC 2119](#) Key Words", [BCP 14](#), [RFC 8174](#), DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, [RFC 8200](#), DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8899] Fairhurst, G., Jones, T., Tüxen, M., Rüngeler, I., and T. Völker, "Packetization Layer Path MTU Discovery for Datagram Transports", [RFC 8899](#), DOI 10.17487/RFC8899, September 2020, <<https://www.rfc-editor.org/info/rfc8899>>.

Authors' Addresses

Guofeng Qian
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China
Email: Qiangguofeng@huawei.com

Tianran Zhou
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China
Email: Zhoutianran@huawei.com