Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2019

R. Hinden
Check Point Software
G. Fairhurst
University of Aberdeen

October 16, 2018

IPv6 Minimum Path MTU Hop-by-Hop Option
draft-hinden-6man-mtu-option-00

Abstract

This document specifies a new Hop-by-Hop IPv6 option that is used to record the minimum Path MTU from a source to a destination host. This collects a minimum recorded MTU along the path to the destination. The value can then be communicated back to the source host by an ICMPv6 Packet Too Big message.

This Hop-by-Hop option is intended to be used in environments like Data Centers and on paths between Data Centers, to allow them to better take advantage of paths able to support a large Path MTU.

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 April 19, 2019.

Copyright Notice

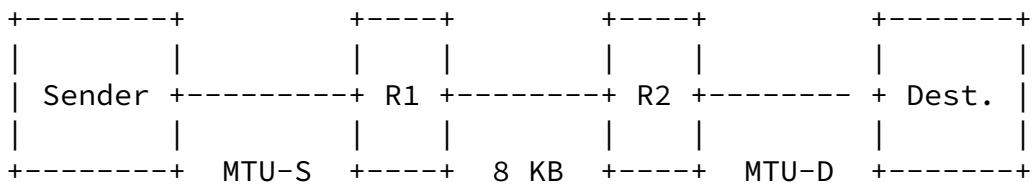
Table of Contents

1. Introduction

This draft proposes a new Hop-by-Hop Option to be used to record the minimum MTU along a path between the source and destination nodes. The source node creates a packet with this Hop-by-Hop Option and fills the Reported PMTU Field in the option with the MTU of the outbound link that will be used to forward the the packet towards the destination.

At each subsequent hop where the option is processed, the router compares the value of the Reported PMTU in the option and the MTU of its outgoing link. If the MTU of the outgoing link is less than the Reported PMTU specified in the option, it rewrites the value in the Option Data with the smaller value. When the packet arrives at the Destination node, the Destination node can send the minimum reported PMTU value back to the Source Node. This can be done by creating an ICMPv6 Packet Too Big message.

The figure below can be used to illustrate the operation of the method. In this case, the path between the Sender and Destination nodes comprises three links, the sender has a link MTU of size MTU-S, the link between routers R1 and R2 has an MTU of size 8 KBytes, and the final link to the destination has an MTU of size MTU-D.



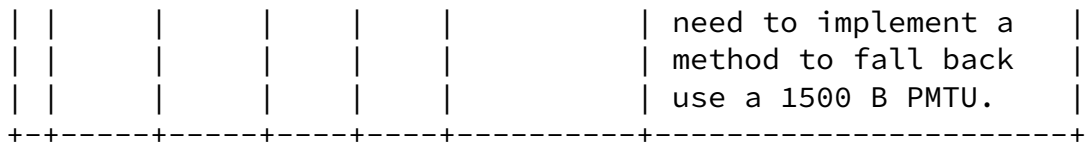
The scenarios are described:

Scenario 1, considers all links to have an 8 KByte MTU and the method is supported by both routers.

Scenario 2, considers the destination link to have an MTU of 1500 Byte. This is the smallest MTU, router R2 resets the reported PMTU to 1500 Byte and this is detected by the method. Had there been another smaller MTU at a link further along the path that supports the method, the lower PMTU would also have been detected.

Scenario 3, considers the case where the router preceding the smallest link does not support the method, and the method then fails to detect the actual PMTU. These scenarios are summarized in the table below. This scenario would also arise if the PTB message was not delivered to the sender.

	MTU-S	MTU-D	R1	R2	Rec PMTU	Note
1	8KB	8KB	H	H	8 KB	Endpoints attempt to use an 8 KB PMTU.
2	8KB	1500B	H	H	1500 B	Endpoints attempt to use a 1500 B PMTU.
3	8KB	1500B	H	-	8 KB	Endpoints attempt to use an 8 KB PMTU, but



IPv6 as specified in [RFC8200] allows nodes to optionally process Hop-by-Hop headers. Specifically from Section 4:

- The Hop-by-Hop Options header is not inserted or deleted, but may be examined or processed by any node along a packet's delivery path, until the packet reaches the node (or each of the set of

nodes, in the case of multicast) identified in the Destination Address field of the IPv6 header. The Hop-by-Hop Options header, when present, must immediately follow the IPv6 header. Its presence is indicated by the value zero in the Next Header field of the IPv6 header.

- NOTE: While [RFC2460] required that all nodes must examine and process the Hop-by-Hop Options header, it is now expected that nodes along a packet's delivery path only examine and process the Hop-by-Hop Options header if explicitly configured to do so.

The Hop-by-Hop Option defined in this document is designed to take advantage of this property of how Hop-by-Hop options are processed. Nodes that do not support this Option SHOULD ignore them. This can mean that the value returned in the response message does not account for all links along a path.

2. 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.

3. Applicability Statements

This Hop-by-Hop Option header is intended to be used in environments such as Data Centers and on paths between Data Centers, to allow them

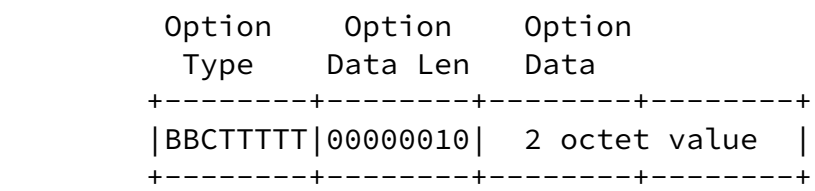
to better take advantage of a path able to support a large PMTU. For example, it helps inform a sender that the path includes links that have a MTU of 9,000 Bytes. This has many performance advantages compared to the current practice of limiting packets to 1280 Bytes.

The design of the option is sufficiently simple that it could be executed on a router's fast path. To create critical mass for this to happen will have to be a strong pull from router vendors customers. This could be the case for connections within and between Data Centers.

The method could also be useful in other environments, including the general Internet.

4. IPv6 Minimum Path MTU Hop-by-Hop Option

The Minimum Path MTU Hop-by-Hop Option has the following format:



Option Type:

BB 00 Skip over this option and continue processing.

C 1 Option data can change en route to the packet's final destination.

TTTTT [Option Type to Be Assigned by IANA]

Length: 2 Note the size of the Option Data field supports Path MTU values from 0 to 65,535 octets.

Value: n The Reported PMTU in octets, reflecting the smallest link MTU that the packet experienced across the path.

5. Router, Host, and Transport Behaviors

5.1. Router Behaviour

Routers that do not support Hop-by-Hop options SHOULD ignore this option and forward the packet.

Routers that support Hop-by-Hop Options, but do not recognize this option SHOULD ignore it and forward the packet.

Routers that recognize this option SHOULD compare the MTU in the Option Value field and the MTU of the outgoing link. If the MTU of the outgoing link is less than the MTU in the option, the router rewrites the value in the Option Value field with the smaller value.

Discussion:

- The design of this Hop-by-Hop Option makes it feasible to be in the fast path of a router, because the required processing is simple.

5.2. Host Behavior

The source host that supports this option SHOULD create the packet with this Hop-by-Hop Option and fill the reported PMTU field of the option with the MTU of the link field that it will send the packet over on the next hop towards the destination.

Discussion:

- This option need not be sent in all packets when using a Transport protocol.

- In the case of TCP, it could be included in packets carrying a SYN segment as part of the connection set up, or can periodically sent in packets carrying other segments. Including this option in a large packet is not likely to be useful, since the large packet might itself also be dropped by a link along the path with a smaller MTU, preventing the Reported PMTU information from reaching the Destination node.

- The use Transport protocols like UDP are harder to characterize because they range from very short-lived exchanges, to longer

exchanges of packets between the Source and Destination nodes.

- For applications that use Anycast, this option should be included in all packets as the actual destination will vary due to the nature of Anycast.

- To optimise for simple-exchange protocols that only send one or a few packets per transaction, a node could assume the Path MTU is symmetrical, that is where the Path MTU is the same in both directions, or at least not smaller in the return path. This optimisation does not hold when the paths are not symmetric.

- The use of this with DNS and DNSSEC over UDP ought to work as long as the paths are symmetric. The DNS server will learn the Path MTU from the DNS query messages. If the return Path MTU is smaller, then the large DNSSEC response may be dropped and the known problems with PMTUD will occur. DNS and DNSSEC over transport protocols that can carry the Path MTU should work.

A Destination Host MUST NOT respond to each packet received with the option, when the option also carries the same received value. This is necessary to avoid generating excessive feedback traffic. When sending an ICMPv6 Packet Too Big message the node MUST follow the procedures in [RFC4443] and [RFC8201] in order to not send too many ICMPv6 Packet Too Big Messages to the source.

When a Destination Host, that supports this option, receives a packet with this option, it SHOULD first compare the Reported PMTU value with a value received earlier from this source. If this is the first value, or if the received value is lower, it SHOULD record the value as the Received PMTU for the Source of the Packet, and it SHOULD send the new value back to the Source of the packet. This can be done by creating an ICMPv6 Packet Too Big message.

NOTE: The Received PMTU could also be reset by a timer to allow periodic refresh of the state. This would also allow a sender to discover cases where the Path MTU has increased.

Discussion:

- A simple mechanism could only send an ICMPv6 Packet Too Big message the first time this option is received or when the

Received PMTU is reduced. This is good because it limits the number sent, but there is no provision for retransmission of the Path MTU if the ICMPv6 Packet Too Big Message fails to reach the sender, or the sender looses state.

- The Reported PMTU value could increase or decrease over time. For instance, it would increase when the path changes and the packets become then forwarded over a link with a MTU larger than the link previously used.

5.3. Transport Behavior

A transport endpoint using this option needs to use a method to verify the information provided by this option.

The Received PMTU does not necessarily reflect the actual PMTU between the sender and destination. Care therefore needs to be exercised in using this value at the sender. Specifically:

- If the Received PMTU value returned by the Destination is the same as the initial Reported PMTU value, there could still be a router or layer 2 device on the path that does not support this PMTU.

- If the Received PMTU value returned by the Destination is smaller than the initial Reported PMTU value, there is at least one router in the path with a smaller MTU. There could still be another router or layer 2 device on the path that does not support this MTU.

- If the Received PMTU value returned by the Destination is larger than the initial Reported PMTU value, this may be a corrupted, delayed or misordered response, and SHOULD be ignored.

A sender needs to discriminate between the Received PMTU value in a PTB message generated in response to a Hop-by-Hop option requesting this, and a PTB message received from a router on the path.

A PMTUD or PLPMTUD method could use the Received PMTU value as an initial target size to probe the path. This can significantly decrease the number of probe attempts (and hence time taken) to

arrive at a workable PMTU. It has the potential to complete

discovery of the correct value in a single RTT, even over paths that may have successive links configured with lower MTUs.

Since the method can delay notification of an increase in the actual PMTU, the sender SHOULD continue to probe for a PMTU value that is larger than the Received PMTU value.

Since the option consumes less capacity than an a full probe packet, there may be advantage in using this to detect a change in the path characteristics.

Note: Further details to be included in next version.

NOTE: A future version of the document will consider more the impact of ECMP. Specifically, whether a Received PMTU value is maintained by the method for each transport endpoint, or each network address, and how these are best used by methods such as PLPMTUD.

6. IANA Considerations

IANA is requested to assign the new IPv6 Hop-by-Hop Option.

IANA is also requested to register this option in the "Destination Options and Hop-by-Hop Options Registry" [IANA-HBH].

7. Security Considerations

A sender MUST check the quoted packet within the PTB message to validate that the message is in response to a packet that was originated by the sender. Messages that fail this check MAY be logged but the information they contain MUST be discarded.

The method has no way to protect the destination from off-path attack with packets that do not originate from the source. This could be used to inflate or reduce the size of the reported PMTU.

The method solicits a response from the destination, which should be used to force generation of a response. A malicious device could advertise a change size of MTU creating work at the destination, and potentially traffic on the return path to the sender.

TBD

8. Acknowledgments

Helpful comments were received from [your name here] and other members of the 6MAN working group.

9. Change log [RFC Editor: Please remove]

draft-hinden-6man-mtu-option-00, 2018-Oct-16:

Initial draft.

10. References

10.1. Normative References

[IANA-HBH] "Destination Options and Hop-by-Hop Options", <https://www.iana.org/assignments/ipv6-parameters/ipv6-parameters.xhtml#ipv6-parameters-2>.

[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>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[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>.

[RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <https://www.rfc-editor.org/info/rfc8201>.

10.2. Informative References

[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>.

Authors' Addresses

Robert M. Hinden
Check Point Software
959 Skyway Road
San Carlos, CA 94070
USA

Email: bob.hinden@gmail.com

Godred Fairhurst
University of Aberdeen
School of Engineering
Fraser Noble Building
Aberdeen AB24 3UE
UK

Email: gorry@erg.abdn.ac.uk