Destination Originates Internet Control Message Protocol (ICMP) Packet Too Big (PTB) Messages
draft-leddy-6man-truncate-03

Abstract

   This document defines procedures that enhance Path MTU Discovery
   (PMTUD), so that it no longer relies on the network's ability to
   deliver an ICMP Packet Too Big (PTB) message from a downstream router
   to an IPv6 source node. According to these procedures, selected
   packets carry a new IPv6 Destination option. When a downstream
   router cannot forward one of these packets because of MTU issues, it
   truncates the packet, marks it to indicate that it has been
   truncated, and forwards it towards the destination node.

   When the destination node receives a packet that has been truncated,
   it sends an ICMP PTB message to the source node. The source node
   uses MTU information contained by the ICMP PTB message to update its
   PMTU estimate.

   The destination node also examines the new Destination option to
   determine whether it should discard the truncated packet or deliver
   it to an upper-layer protocol.

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 December 13, 2018.

Table of Contents

1. Introduction

   An Internet path connects a source node to a destination node. A
   path can contain links and routers.

   Each link is constrained by the number of bytes that it can convey in
   a single IP packet. This constraint is called the link Maximum
   Transmission Unit (MTU). IPv6 [RFC8200] requires every link to have

an MTU of 1280 bytes or greater. This value is called IPv6 minimum
link MTU.

Likewise, each Internet path is constrained by the number of bytes
that it can convey in a IP single packet. This constraint is called
the Path MTU (PMTU). For any given path, the PMTU is equal to the
smallest of its link MTUs.

IPv6 allows fragmentation at the source node only. If an IPv6 source
node sends a packet whose length exceeds the destination PMTU, the
packet will be discarded. In order to avoid this kind of packet
loss, IPv6 nodes can either:

o Refrain from sending packets whose length exceeds the IPv6 minimum
   link MTU.

o Maintain a running estimate of the destination PMTU and refrain
   from sending packets whose length exceeds that estimate.

IPv6 nodes can execute Path MTU Discovery (PMTUD) [RFC8201]
procedures in order to maintain a running estimate of the destination
PMTU. According to these procedures, the source node produces an
initial PMTU estimate. This initial estimate is equal to the MTU of
the first link along the path to the destination node. It can be
greater than the actual PMTU.

Having produced an initial PMTU estimate, the source node sends
packets to the destination node. If one of these packets is larger
than the actual PMTU, a downstream router will not be able to forward
the packet through the next link along the path. Therefore, the
downstream router discards the packet and sends an Internet Control
Message Protocol (ICMP) [RFC4443] Packet Too Big (PTB) message to the
source node. The ICMP PTB message indicates the MTU of the link
through which the packet could not be forwarded. The source node
uses this information to refine its PMTU estimate.

PMTUD relies on the network's ability to deliver ICMP PTB messages
from the downstream router to the source node. If the network cannot
deliver these messages, a persistent black hole can develop. In this
scenario, the source node sends a packet whose length exceeds the
destination PMTU. A downstream router discards the packet and sends
an ICMP PTB message to the source. However, the network cannot
deliver the ICMP PTB message to the source. Therefore, the source
node does not update its PMTU estimate and it continues to send
packets whose length exceeds the destination PMTU. The downstream
router discards these packets and sends ICMP PTB messages to the
source. These ICMP PTB messages are lost, exactly as previous ICMP
PTB messages were lost.

In some operational scenarios ([Section 3](#)), networks cannot deliver
ICMP PTB messages from a downstream router to the source node.
Therefore, enhanced procedures are required.

This document defines procedures that enhance PMTUD, so that it no
longer relies on the network's ability to deliver an ICMP PTB message
from a downstream router to an IPv6 source node. According to these
procedures, selected packets carry a new IPv6 Destination option.
When a downstream router cannot forward one of these packets because
of MTU issues, it truncates the packet, marks it to indicate that it
has been truncated, and forwards it towards the destination node.

When the destination node receives a packet that has been truncated,
it sends an ICMP PTB message to the source node. The source node
uses MTU information contained by the ICMP PTB message to update its
PMTU estimate.

The destination node also examines the new Destination option to
determine whether it should discard the truncated packet or deliver
it to an upper-layer protocol. If the truncated packet is delivered
to an upper-layer protocol, the upper-layer protocol can infer the
PMTU between the source node and itself from the packet's length.
Having inferred the PMTU, the upper-layer protocol can negotiate a
maximum packet size with its upper-layer peer, thus reducing its
dependence upon PMTUD and IPv6 fragmentation.

While packet truncation may facilitate new upper-layer procedures,
upper-layer procedures are beyond the scope of this document.

[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](#). Operational Considerations

The packet truncation procedures described herein make PMTUD more
resilient when:

o The network can deliver ICMP PTB messages from the destination
   node to the source node.

o The network cannot deliver ICMP PTB messages from a downstream
   router to the source node.

The following are operational scenarios in which packet truncation
procedures can make PMTUD more resilient:

o The destination node has a viable route to the source node, but
   the downstream router does not.

o The source node is protected by a firewall that administratively
   blocks all packets except for those from specified subnetworks.
   The destination node resides in one of the specified subnetworks,
   but the downstream router does not.

o The source address of the original packet (i.e., the packet that
   elicited the ICMP PTB message) was an anycast address. Therefore,
   the destination address of the ICMP PTB message is the same
   anycast address. In this case, an ICMP PTB message from the
   destination node is likely to be delivered to the correct anycast
   instance. By contrast, an ICMP PTB message from a downstream
   router is less likely to be delivered to the correct anycast
   instance.

Packet truncation procedures do not make PMTUD more resilient when
the network cannot reliably deliver any ICMP PTB messages to the
source node. The following are operational scenarios where the
network cannot reliably deliver any ICMP PTB messages to the source
node:

o The source node is protected by a firewall that administratively
   blocks all ICMP PTB messages.

o The source node is an anycast instance served by a load-balancer
   as defined in [RFC7690]. The load-balancer does not implement the
   mitigations defined in [RFC7690].

4. Reference Topology

```
    -------------          -------------          -------------
    |   Upper   |          |           |          |   Upper   |
    |   Layer   |          |           |          |   Layer   |
    |           |          |           |          |           |
    |    IP     |<-------->|    IP     |<-------->|    IP     |
    |   Layer   |   MTU    |   Layer   |   MTU    |   Layer   |
    -------------   9000   -------------   1500   -------------
       Source                 Router             Destination
        Node                                         Node
```

                            Figure 1

Figure 1 depicts a reference topology that is used in this document.
The reference topology includes a source node, a router, and a
destination node. The link that connects the source node to the
router has an MTU of 9000 bytes. The link that connects the router
to the destination node has an MTU of 1500 bytes. Therefore, the
PMTU between the source node and the destination node is 1500 bytes.

5. Truncation Option

```
       0                   1                   2
       0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |  Option Type  |  Opt Data Len |T|D|R| Reserved|
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                            Figure 2

Figure 2 depicts the Truncation option. The IPv6 Destination Options
header MAY include the Truncation option. The IPv6 Hop-by-hop header
MUST NOT include the Truncation option.

o Option Type - Truncation option. Value TBD by IANA. See Notes
   below.

o Opt Data Len - Length of Option Data, measured in bytes. MUST be
   equal to 1.

o T - Indicates that the packet has been truncated. If the packet
   has not been truncated, it SHOULD be delivered to an upper-layer
   protocol.

o D - Indicates that the packet SHOULD be delivered to an upper-
   layer protocol even if it has been truncated. If the D-bit is set
   to zero, the packet MUST NOT be delivered to an upper-layer
   protocol if it has been truncated.

o R - Request to deliver truncated packets.

A packet MUST NOT contain multiple instances of the Truncation
option. If a node receives a packet that contains multiple instances
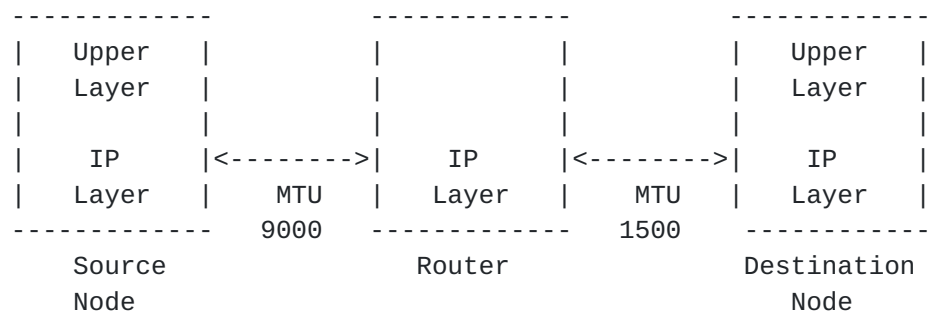of the Truncation option, it processes the first instance and ignores
all subsequent instances.

NOTE 1: The highest-order two bits of the Option Type (i.e., the
"act" bits) are 10. These bits specify the action taken by a
destination node that does not recognize Truncation option. The
required action is to discard the packet and, regardless of whether
or not the packet's Destination Address was a multicast address, send

an ICMP Parameter Problem, Code 2, message to the packet's Source
Address, pointing to the unrecognized Option Type.

NOTE 2: The third highest-order bit of the Option Type (i.e., the
"chg" bit) is 1. This indicates that Option Data can be modified
along the path between the packet's source and its destination.

6. PMTU Signaling Procedures

In the Reference Topology (Figure 1), the upper-layer protocol that
resides on the source node submits a packet to its local IP layer.
The packet includes a Truncation option (Section 5). Within the
Truncation option:

o The T-bit is set to zero, indicating that the packet has not been
   truncated.

o The D-bit is set to zero, indicating that the packet MUST NOT be
   delivered to an upper-layer protocol if the packet has been
   truncated.

o The R-bit is set to zero, indicating that the source node does not
   request delivery of truncated packets

The packet length is 500 bytes. Because the packet length is less
than the destination PMTU, the packet can be delivered without
encountering MTU issues.

The IP layer on the source node forwards the packet to the downstream
router and the downstream router forwards the packet to the
destination node. The IP layer on the destination node examines the
Destination Option header. Because it recognizes the Truncation
option, and because the packet has not been truncated, it delivers
the packet to an upper-layer protocol.

Now, the upper-layer protocol that resides on the source node submits
another packet to its local IP layer. This packet is identical to
the first, except that the packet length is 2000 bytes. Because the
packet length is greater than the destination PMTU, the packet cannot
be delivered without encountering MTU issues.

The IP layer on the source node forwards the packet to the downstream
router but the downstream router cannot forward the packet because
its length exceeds the MTU of the next-hop link. Because an MTU
issue has been encountered, the IP Layer on the downstream router
examines the Destination Options header, searching for a Truncation
option. (Normally, the downstream router would ignore the
Destination Options header).

Because the downstream router finds and recognizes the Truncation option, it:

o Truncates the packet, so that its new length equals the MTU of the next-hop link.

o Updates the Payload Length field in the IPv6 header as appropriate.

o Sets the T-bit in the Truncation option.

o Forwards the packet to the destination node.

The IP layer on the destination node examines the Destination Option header. Because it recognizes the Truncation option, and because the packet has been truncated, it sends an ICMP PTB message to the source node. The MTU field in the ICMP PTB message is set to the packet's length.

The IP layer on the destination node then discards the packet because the packet has been truncated and the D-bit is set to 0. It does not deliver the packet to an upper-layer protocol.

As per [RFC8201], the source node updates its PMTU estimate using information contained by the ICMP PTB message.

7. Truncation Considerations

A packet can be truncated multiple times.

```
-------------            -------------            -------------
-------------
|   IP     |<-------->|    IP      |<-------->|    IP      |<-------->|
IP     |
|  Layer   |   MTU    |  Layer     |   MTU    |  Layer     |   MTU     |
Layer    |
-------------   9000    -------------   4000   -------------   1500
-------------
    Source                 Router1                Router 2
Destination
     Node                                                            Node
```
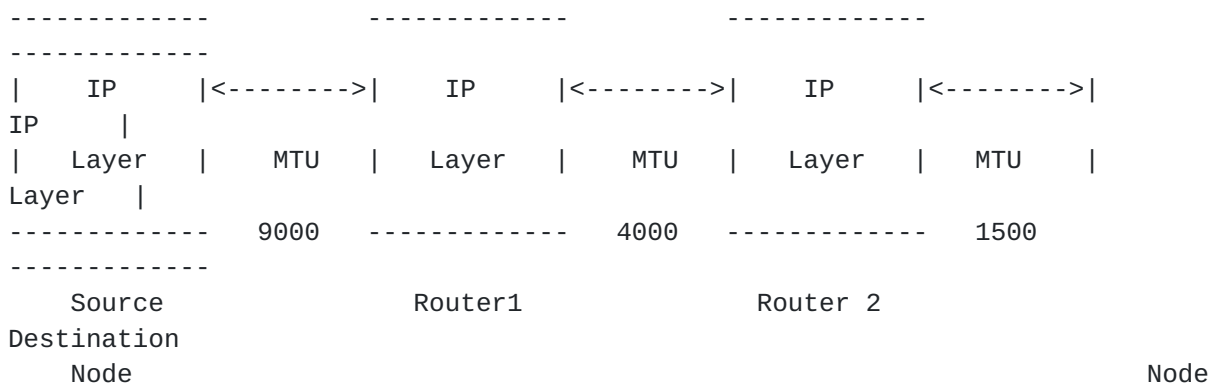
Figure 3: Double Truncation

Figure 3 depicts a network that contains a Source Node, Router1, Router2 and a Destination Node. The link that connects the Source Node to Router1 has an MTU of 9000 bytes. The link that connects Router1 to Router2 has an MTU of 4000 bytes, and the link that

connects Router2 to the Destination Node has an MTU of 1500 bytes.

Assume that the Source Node sends a packet to the Destination Node.
The packet is 8000 bytes long. When Router1 receives this packet, it

identifies the next-hop towards the destination. This is the link
that connects Router1 to Router2. Router 1 encounters an MTU issue,
because the packet length (8000 bytes) is greater than the MTU
associated with the next-hop link (4000 bytes). Therefore, Router 1
truncates the packet to 4000 bytes, sets the T-bit in the Truncation
Option, and forwards the packet towards Router2. When Router2
receives this packet, it identifies the next-hop towards the
destination. This is the link that connects Router2 to the
Destination Node. Router 2 encounters an MTU issue, because the
packet length (4000 bytes) is greater than the MTU associated with
the next-hop link (1500 bytes). Therefore, Router 2 truncates the
packet to 1500 bytes, sets the T-bit in the Truncation Option, and
forwards the packet towards the Destination Node. The Destination
Node sends an ICMP PTB packet to the source node. The MTU field in
the ICMP PTB field is set to 1500.

A truncated packet MUST contain the basic IPv6 header, all extension
headers and the first upper-layer header. When a router cannot
forward a packet through the next-hop link due to MTU issues, and the
total length of the basic IPv6 header, all extension headers, and
first upper-layer header exceeds the MTU of the next-hop link, the
router MUST discard the packet and send and ICMP PTB message to the
source.

Source nodes MUST NOT emit packets that contain both the Fragment
Header and Truncation Option.

Routers MUST NOT truncate packets that include the Fragment header.
When a router cannot forward a packet through the next-hop link due
to MTU issues, and the packet includes a Fragment header, the router
MUST discard the packet and send and ICMP PTB message to the source.

Routers MUST NOT emit truncated packets whose length is less than the
IPv6 minimum link MTU.

8. ICMP Considerations

When a destination node receives a truncated packet whose length is
less than the IPv6 minimum link MTU, the destination node MUST
discard the packet. It MUST NOT send an ICMP PTB message to the
packet's source and it MUST NOT deliver the packet to an upper-layer
protocol.

9. Delivering Truncated Packets

A destination node MUST NOT deliver a truncated packet to an upper-
layer protocol if the D-bit is set to zero. However, a destination

node SHOULD deliver a truncated packet to an upper-layer protocol if
the D-bit is set to one.

The upper-layer protocol on the source node determines the D-bit
value. The following are possible behaviors:

o Some upper-layer protocols never set the D-bit.

o Some upper-layer protocols always set the D-bit.

o Some upper-layer protocols only set the D-bit when requested to do
   so by the upper-layer protocol on the destination node. These
   upper-layer protocols interpret the receipt of a Truncation option
   with the R-bit set as a request to set the D-bit on all subsequent
   packets.

10. Backward Compatibility

The procedures described in Section 6 of this document assume that
the source node, downstream router and destination node all
recognized the Truncation option. This section explores backwards
compatibility, where one or more nodes do not recognize the
Truncation option.

If the destination node does not recognize the Truncation option, and
it receives a packet that includes the Truncation option, it discards
the packet and, regardless of whether or not the packet's Destination
Address was a multicast address, sends an ICMP Parameter Problem,
Code 2, message to the packet's Source Address, pointing to the
unrecognized Option Type. This behavior is determined by the
highest-order two bytes of the Option Type. When the source node
receives the ICMP Parameter Problem message, it refrains from sending
packets that contain the Truncation option.

If the downstream router does not recognize the Truncation option and
it receives a packet that contains the Truncation option and that
packets length does not exceed the next-hop MTU, the downstream
router forwards the packet, without examining the Truncation option
or any other Destination option. If the downstream router does not
recognize the Truncation option and it receives a packet that
contains the Truncation option and that packets length exceeds the
next-hop MTU, the downstream discards the packet and sends an ICMP
PTB message to the source node, as per [RFC8200].

In all cases mentioned above, PMTUD continues to function as
specified in [RFC8201].

11. Optimizations

The procedures described in Section 6 of this document can be
optimized by omitting the Truncation option on packets whose length
is known to be less than the destination PMTU (e.g., packets whose
length is less than the IPv6 minimum link MTU).

12. Upper-Layer Considerations

The procedures described herein rely upon:

o The network's ability to convey packets that contain destination
   options from a source node to a destination node.

o The network's ability to convey ICMP Parameter Problem messages in
   the reverse direction.

Operational experience [RFC7872] reveals that a significant number of
networks drop packets that contain IPv6 destination options.
Likewise, many networks drop ICMP Parameter Problem messages.

[I-D.bonica-6man-unrecognized-opt] describes procedures that upper-
layer protocols can execute to verify that the above-mentioned
requirements are satisfied. Upper-layer protocols can execute these
procedures before emitting packets that contain the Truncation
option.

13. Encapsulating Security Payload Considerations

An IPv6 packet can contain both:

o An Encapsulating Security Payload (ESP) [RFC4303] header.

o The Truncation Option.

In this case, the packet MUST contain a Destination Options header
that precedes the ESP. That Destination Options header contains the
Truncation Option and is not protected by the ESP. The packet MAY
also contain another Destination Options header the follows the ESP.
This Destination Options header is protected by the ESP and MUST NOT
contain the Truncation Option.

As per RFC 4303, a packet can contain two Destination Option headers
one preceding the ESP and one following the ESP.

14. **Extension Header Considerations**

   According to [RFC8201], the following IPv6 extension headers can
   carry options:

   o The Hop-by-hop Options header.

   o The Destination Options header.

   The Hop-by-hop Options header is examined by the destination node.
   It can also be examined by intermediate nodes (i.e., nodes along the
   path between the source and the destination), so long as those nodes
   are configured to process Hop-by-hop options. By contrast, the
   Destination Options header is examined by the destination node only.

   The Truncation option is examined by:

   o The destination node.

   o Intermediate nodes, but only on an exception basis (i.e., when the
      intermediate node cannot forward the packet due to MTU issues)

   If performance were not a concern, the Hop-by-hop Options header
   could carry the Truncation Option. The destination node would
   examine the Truncation option, as would every intermediate node.
   However, the performance impact would not be acceptable.

   By contrast, the Destination Option can carry the truncation option,
   so long as intermediate nodes can examine the Destination Option
   header on an exception basis (e.g., when the packet cannot be
   forwarded due to MTU issues). [RFC2473] sets a precedent for
   intermediate nodes examining the Destination Options header on an
   exception basis. (See the Tunnel Encapsulation Limit.)

   Therefore, The IPv6 Destination Options header MAY include the
   Truncation option and the IPv6 Hop-by-hop header MUST NOT include the
   Truncation option.

15. **Security Considerations**

   PMTUD is vulnerable to ICMP PTB forgery attacks. The procedures
   described herein do nothing to mitigate that vulnerability.

   The procedures described herein are susceptible to a new variation on
   that attack, in which an attacker forges a truncated packet. In this
   case, the attackers cause the destination node to produce an ICMP PTB
   message on their behalf. To some degree, this vulnerability is
   mitigated, because the destination node will not emit an ICMP PTB

message in response to a truncated packet whose length is less than
the IPv6 minimum link MTU.

16. IANA Considerations

IANA is requested to allocate a codepoint from the Destination
Options and Hop-by-hop Options registry
(https://www.iana.org/assignments/ipv6-parameters/
ipv6-parameters.xhtml#ipv6-parameters-2). This option is called
"Truncation". The "act" bits are 10 and the "chg" bit is 1.

17. Acknowledgements

Special thanks to Mike Heard, Geoff Huston, Joel Jaegglii, Andy
Smith, and Jinmei Tatuya who reviewed and commented on this document.

18. References

18.1. Normative References

[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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing
           Architecture", RFC 4291, DOI 10.17487/RFC4291, February
           2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)",
           RFC 4303, DOI 10.17487/RFC4303, December 2005,
           <https://www.rfc-editor.org/info/rfc4303>.

[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>.

18.2. Informative References

   [I-D.bonica-6man-unrecognized-opt]
               Bonica, R. and J. Leddy, "The IPv6 Unrecognized Option",
               draft-bonica-6man-unrecognized-opt-00 (work in progress),
               June 2018.

   [RFC2473]   Conta, A. and S. Deering, "Generic Packet Tunneling in
               IPv6 Specification", RFC 2473, DOI 10.17487/RFC2473,
               December 1998, <https://www.rfc-editor.org/info/rfc2473>.

   [RFC7690]   Byerly, M., Hite, M., and J. Jaeggli, "Close Encounters of
               the ICMP Type 2 Kind (Near Misses with ICMPv6 Packet Too
               Big (PTB))", RFC 7690, DOI 10.17487/RFC7690, January 2016,
               <https://www.rfc-editor.org/info/rfc7690>.

   [RFC7872]   Gont, F., Linkova, J., Chown, T., and W. Liu,
               "Observations on the Dropping of Packets with IPv6
               Extension Headers in the Real World", RFC 7872,
               DOI 10.17487/RFC7872, June 2016,
               <https://www.rfc-editor.org/info/rfc7872>.

Authors' Addresses

   John Leddy
   Comcast
   1717 John F Kennedy Blvd.
   Philadelphia, PA 19103
   USA

   Email: john_leddy@comcast.com

   Ron Bonica
   Juniper Networks
   2251 Corporate Park Drive
   Herndon, Virginia 20171
   USA

   Email: rbonica@juniper.net