6man
Internet-Draft
Intended status: Standards Track
Expires: December 31, 2018

J. Leddy
Comcast
R. Bonica
Juniper Networks
June 29, 2018

IPv6 Packet Truncation
draft-leddy-6man-truncate-04

Abstract

This document defines IPv6 packet truncation procedures. When an IPv6 source node originates a packet, it can mark the packet as being eligible for truncation and forward it towards its destination. If an intermediate node cannot forward the packet because of an MTU issue, it truncates the packet, marks it as being truncated, and, again, forwards it towards its destination. When the destination node receives the packet, it detects that it has been truncated and sends an ICMP message to the source node. The ICMP message contains MTU information that the source node uses to update its Path MTU estimate.

The above-mentioned procedures enhance Path MTU Discovery (PMTUD) by eliminating its reliance on the network's ability to deliver ICMP messages from an intermediate node to the source node. However, the above-mentioned procedures require the network to deliver ICMP messages from the destination node to the source node.

Table of Contents

1. Introduction

An Internet path connects a source node to a destination node. A path can contain links and intermediate nodes (e.g., 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 PMTU, an intermediate node will discard the packet. In order to prevent this, IPv6 nodes can either:

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

- Maintain a running estimate of the 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 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. 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, an intermediate node will not be able to forward the packet through the next link along the path. Therefore, the intermediate node 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 intermediate node 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 PMTU. An intermediate node 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 PMTU. The intermediate node discards these packets and sends more 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 an intermediate node to the source node. Therefore, enhanced procedures are required.

This document defines IPv6 packet truncation procedures. When an IPv6 source node originates a packet, it can mark the packet as being eligible for truncation and forward it towards its destination. If an intermediate node cannot forward the packet because of an MTU issue, it truncates the packet, marks it as being truncated, and, again, forwards it towards its destination. When the destination node receives the packet, it detects that it has been truncated and sends an ICMP message to the source node. The ICMP message contains MTU information that the source node uses to update its Path MTU estimate.

The above-mentioned procedures enhance PMTUD by eliminating its reliance on the network's ability to deliver ICMP messages from an intermediate node to the source node. However, the above-mentioned procedures require the network to deliver ICMP messages from the destination node to the source node.

By default, destination nodes discard truncated packets and do not deliver them to upper-layer protocols. However, upper-layer protocols can register for delivery of truncated packets. When an upper-layer protocol receives a truncated packet, it 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 reliance upon PMTUD and IPv6 fragmentation.

While IPv6 packet truncation may facilitate new upper-layer procedures, upper-layer procedures are beyond the scope of this document. In particular, this document does not address the behavior of upper-layer protocols that register for delivery of truncated packets.

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:

- The network can deliver ICMP messages from the destination node to the source node.
- The network cannot deliver ICMP messages from an intermediate node to the source node.

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

- The destination node has a viable route to the source node, but the intermediate node does not.
- 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 intermediate node does not.
- The source address of the original packet (i.e., the packet that elicited the ICMP message) was an anycast address. Therefore, the destination address of the ICMP message is the same anycast address. In this case, an ICMP message from the destination node is likely to be delivered to the correct anycast instance. By contrast, an ICMP message from an intermediate node 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 messages to the source node. The following are operational scenarios where the network cannot reliably deliver any ICMP PTB messages to the source node:

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

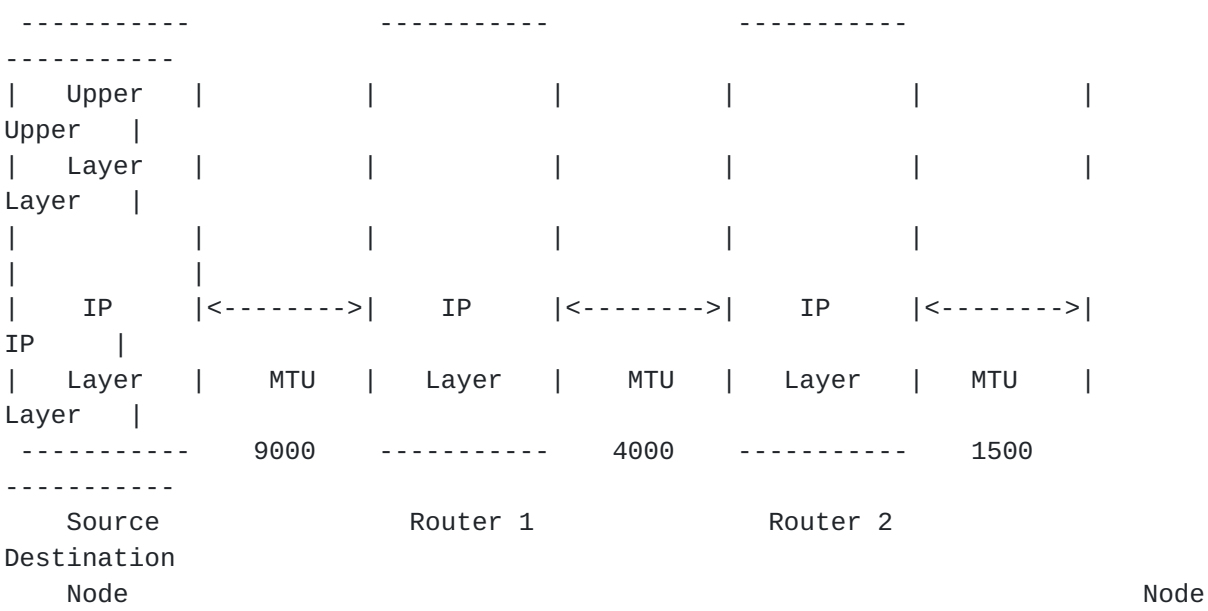

Figure 1: Reference Topology

Figure 1 depicts a network that contains a Source Node, intermediate nodes (i.e., Router 1, Router 2), and a Destination Node. The link that connects the Source Node to Router 1 has an MTU of 9000 bytes. The link that connects Router 1 to Router 2 has an MTU of 4000 bytes, and the link that connects Router 2 to the Destination Node has an MTU of 1500 bytes. The PMTU between the Source Node and the Destination Node is 1500 bytes.

This topology will be used in examples throughout the document.

5. New IPv6 Destination Options

This document defines the IPv6 Truncation Eligible option and the IPv6 Truncated Packet option.

5.1. The IPv6 Truncation Eligible Option

The IPv6 Truncation Eligible Option indicates that the packet is eligible for truncation but has not been truncated. It contains the following fields:

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

- Opt Data Len - Length of Option Data, measured in bytes. MUST be equal to 0.

The IPv6 Destination Options header:

- MAY include a single instance of the Truncation Eligible option.

- SHOULD NOT include multiple instances of the Truncation Eligible option.

- SHOULD NOT include both the Truncation Eligible option and the Truncated Packet option.

The IPv6 Hop-by-hop Options header SHOULD NOT include the Truncation Eligible option.

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

NOTE 1: According to [RFC8200], the highest-order two bits of the Option Type (i.e., the "act" bits) specify the action taken by a destination node that does not recognize Option Type. The required action is skip over this option and continue processing the header. Therefore, IANA is requested to assign this Option Type with "act" bits "00".

NOTE 2: According to [RFC8200], the third-highest-order bit (i.e., the "chg" bit) of the Option Type specifies whether or not the Option Data of that option can change en route to the packet's final destination. Because this option contains no Option Data, IANA can assign this Option Type without regard to the "chg" bit.

5.2. The IPv6 Truncated Packet Option

The IPv6 Truncated Packet Option indicates that the packet has been truncated and is eligible for further truncation. It contains the following fields:

- Option Type - Truncated Packet option. Value TBD by IANA. See Notes below.

- Opt Data Len - Length of Option Data, measured in bytes. MUST be equal to 0.

The IPv6 Destination Options:

- MAY include a single instance of the Truncated Packet option.

- SHOULD NOT include multiple instances of the Truncated Packet option.

- SHOULD NOT include both the Truncated Packet option and the Truncation Eligible option.

The IPv6 Hop-by-hop Options header SHOULD NOT include the Truncated Packet option.

Source nodes MUST NOT emit packets that contain both the Fragment Header and Truncated Packet option.

NOTE 1: According to [RFC8200], the highest-order two bits of the Option Type (i.e., the "act" bits) specify the action taken by a destination node that does not recognize Option Type. 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. Therefore, IANA is requested to assign this Option Type with "act" bits "10".

NOTE 2: According to [RFC8200], the third-highest-order bit (i.e., the "chg" bit) of the Option Type specifies whether or not the Option Data of that option can change en route to the packet's final destination. Because this option contains no Option Data, IANA can assign this Option Type without regard to the "chg" bit.

6. PMTU Signaling Procedures

In the Reference Topology (Figure 1), an upper-layer protocol that resides on the Source Node causes the IP layer to emit a packet. The packet contains a Destination Options header and the Destination Options header contains a Truncation Eligible option. The total packet length, including all headers and the payload, is 1000 bytes. Because the total packet length is less than the PMTU, the packet can be delivered to the Destination Node without encountering any MTU issues.

The IP layer on the Source Node forwards the packet to the Router 1, Router 1 forwards the packet to Router 2, and the Router 2 forwards the packet to the Destination Node. The IP layer on the Destination Node examines the Destination Options header and finds the Truncation Eligible option. The Truncation Eligible option requires no action by the Destination Node. Therefore, the Destination Node processes the next header and delivers the packet to an upper-layer protocol.

Subsequently, the upper-layer protocol that resides on the Source Node causes the IP layer to emit another packet. This packet is identical to the first, except that the total packet length is 2000 bytes. Because the packet length is greater than the PMTU, this packet cannot be delivered without encountering an MTU issue.

The IP layer on the source node forwards the packet to Router 1. Router 1 forwards the packet to Router 2, but the Router 2 cannot forward the packet because its length exceeds the MTU of the next link in the path. Because an MTU issue has been encountered, Router 2 examines the Destination Options header, searching for either a

Truncation Eligible option or a Truncated Packet option. (Normally, the Router 2 would ignore the Destination Options header).

Because Router 2 finds one of the above-mentioned options, it:

- Truncates the packet, so that its total length equals the MTU of the next link in the path.
- Updates the Payload Length field in the IPv6 header.
- Overwrites all instances of the Truncation Eligible option with a Truncated Packet option.
- Forwards the packet to the Destination Node.

The IP layer on the Destination Node receives the packet and examines the Destination Options header. Because it finds the Truncated Packet option, 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.

By default, the IP layer on the Destination Node discards the truncated packet, without delivering it to any upper-layer protocol. However, the upper-layer protocol can register for the delivery of truncated packets.

When the Source Node receives the ICMP PTB message, it updates its PMTU estimate, as per [RFC8201].

7. Truncation Considerations

A packet can be truncated multiple times. In the Reference Topology (Figure 1), assume that the Source Node sends a 5000 byte packet to the Destination Node. Using the procedures described in Section 6, Router 1 truncates this packet to 4000 bytes and Router 2 truncates it again, to 1500 bytes.

A truncated packet MUST contain the basic IPv6 header, all extension headers and the first upper-layer header. When an intermediate node cannot forward a packet 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 link in the path, the intermediate node MUST discard the packet and send and ICMP PTB message to the source node. It MUST NOT truncate the packet.

A truncated packet MUST NOT include the Fragment header. When an intermediate node cannot forward a packet due to MTU issues, and the packet contains a Fragment header, the intermediate node MUST discard

the packet and send and ICMP PTB message to the source node. It MUST NOT truncate the packet.

A truncated packet must have a total length that is greater than or equal to the IPv6 minimum link MTU.

8. Destination Node Considerations

The following packet types are invalid::

- Packets that contain the Packet Truncated option and the Fragment Header.

- Packets that contain the Packet Truncated option and have a total length less than the IPv6 minimum link MTU.

When the destination node receives an invalid packet, it MUST:

- Discard the packet, without delivering it to any upper-layer protocol, regardless of whether the upper-layer protocol has registered for delivery of truncated packets.

- Send an ICMP Parameter Problem, Code 2, message to the packet's Source Address, pointing to the Truncated Packet option.

9. Backward Compatibility

The procedures described in Section 6 assume that all nodes recognize the Truncation Eligible option and Truncated Packet option. This section explores backwards compatibility scenarios, where one or more nodes do not recognize the above-mentioned options.

Assume that an intermediate node does not recognize the Truncation Eligible option or the Truncated Packet option. When that node receives a packet that it cannot forward because of an MTU issue, its behavior is as described in [RFC8200]. The intermediate node discards the packet and sends and ICMP PTB message to the source node. It does not examine the Destination Options header, searching for the above-mentioned options and it does not truncate the packet.

Now assume that a destination node does not recognize the Truncation Eligible option. When that node receives a packet that contains the Truncation Eligible option, its behavior is determined by the highest-order two bits of the Option Type (i.e., the "act" bits). Because the "act" bits are equal to "00", the destination node skips over the option and continues to process the packet. This is exactly what the destination node would have done if it had recognized the Truncation Eligible option.

Finally, assume that a destination node does not recognize the Truncated Packet option. When that node receives a packet that contains the Truncated Packet option, its behavior is determined by the highest-order two bits of the Option Type (i.e., the "act" bits). Because the "act" bits are equal to "10", the destination node discards 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 Truncated Packet option. The destination node does not emit an ICMP PTB message and it does not deliver the packet to an upper-layer protocol.

The source node takes appropriate action when it receives the ICMP Parameter Problem message.

10. Optimizations

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

11. Upper-Layer Considerations

The procedures described herein rely upon the networks ability:

- To convey packets that contain destination options from the source node to the destination node.

- 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 Eligible option.

12. Encapsulating Security Payload Considerations

An IPv6 packet can contain both:

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

- The Truncation Eligible Option.

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

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

13. Extension Header Considerations

According to [RFC8200], the following IPv6 extension headers can contain options:

- The Hop-by-hop Options header.

- The Destination Options header.

The Hop-by-hop option can be examined by each node along the path to a packet's destination. Destination options are examined by the destination node only. However, [RFC2473] provides a precedent for intermediate nodes examining the Destination options on an exception basis. (See the Tunnel Encapsulation Limit.)

The Truncation Eligible option and the Truncated Packet option are examined by:

- Intermediate nodes, on an exception basis (i.e, when the packet cannot be forwarded due to MTU issues).

- The Destination node.

Therefore, the above-mentioned options can be processed most efficiently when they are contained by the Destination Option header. When contained by the Destination Options header, the above-mentioned options are examined by intermediate nodes on an exception basis, only when they are relevant. If contained by the Hop-by-hop Options header, they are always examined by intermediate nodes, even when they are irrelevant.

14. 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.

15. IANA Considerations

IANA is requested to allocate the following codepoints from the Destination Options and Hop-by-hop Options registry (https://www.iana.org/assignments/ipv6-parameters/ipv6-parameters.xhtml#ipv6-parameters-2).

- Truncation Eligible ("act-bits" are "00. "chg-bit" can be either 0 or 1.)

- Truncated Packet ("act-bits" are "10". "chg-but can be either 0 or 1.)

16. Acknowledgements

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

17. References

17.1. 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>.

[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>.

17.2. Informative References

[I-D.bonica-6man-unrecognized-opt] Bonica, R. and J. Leddy, "The IPv6 Unrecognized Option", draft-bonica-6man-unrecognized-opt-01 (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