IPSECME Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 1, 2018
S. Piriyath
U. Mangla
N. Melam
R. Bonica
Juniper Networks
February 28, 2018

Packetization Layer Path Maximum Transmission Unit Discovery (PLPMTUD) For IPsec Tunnels

draft-spiriyath-ipsecme-dynamic-ipsec-pmtu-01

Abstract

This document describes Packetization Layer PMTU Discovery (PLPMTUD) procedures for IPSec tunnels. In these procedures, the encrypting node discovers and maintains a running estimate of the tunnel MTU. In order to do this, the encrypting nodes sends Probe Packets of various size through the IPSec tunnel. If the size of Probe Packet exceeds the tunnel MTU, a downstream node discards the packet and sends an ICMP PTB message to the encrypting node. The encrypting node ignores the ICMP PTB message. If the size of the Probe Packet does not exceed the tunnel MTU and the decrypting node receives the Probe Packet, the decrypting node sends an Acknowledgement Packet to encrypting node through the IPSec tunnel. The Acknowledgement Packet indicates the size of the Probe Packet.

The procedures described in this document are applicable to IPSec tunnels that are signaled by IKEv2 and provide authentication services.

Copyright Notice

Table of Contents

1. Introduction

IPsec [RFC4301] tunnels provides private and/or authenticated connectivity between an encrypting node and a decrypting node. An IPsec tunnel is constrained by the number of bytes that it can convey in a single packet, without fragmentation of any kind. This constraint is called the tunnel Maximum Transmission Unit (MTU). An IPSec tunnel's MTU can be calculated as its Path MTU (PMTU) minus IPSec tunnel overhead, where:

- o PMTU is the smallest MTU of all the links forming a path between the encrypting node and the decrypting node.
- o IPSec tunnel overhead is the maximum number of bytes required for padding (by the encryption algorithm) plus the number of bytes required for IPSec encapsulation.

When forwarding a packet through an IPSec tunnel, the encrypting node compares the packet's length to the tunnel MTU. If the packet length is less than or equal to the tunnel MTU, the encrypting node encrypts the packet, encapsulates it and forwards it through the IPSec tunnel.

If the packet length is greater than the tunnel MTU and the packet cannot be fragmented, the encrypting node discards the packet and sends an ICMP [RFC0792] [RFC4443] Packet Too Big (PTB) message to the packet's source.

If the packet length is greater than the tunnel MTU and the packet can be fragmented, the encrypting node can execute either of the following procedures:

- Fragment, encrypt and encapsulate (FEE)

- Encrypt, encapsulate and fragment (EEF)

If the encrypting node executes FEE procedures, it fragments the packet first. Then it encrypts, encapsulates and forwards each fragment. When a fragment arrives at the decrypting node, the decrypting node decapsulates and decrypts the fragment. Finally, the decrypting node forwards the fragment to its ultimate destination, where it can be reassembled.

If the encrypting node executes EEF procedures, it encrypts and encapsulates the packet first. Then it fragments the resulting packet and forwards each fragment to the decrypting node. When the decrypting node has received all fragments, it reassembles the packet, decapsulates and decrypts it. Finally, it forwards the packet, in one piece, to its ultimate destination.

In the paragraphs above, IPv4 [RFC0791] packets with the Don't Fragment (DF) bit set to zero can be fragmented. IPv6 [RFC8200] packets and IPv4 packets with the DF bit set to one cannot be fragmented.

In the above-described procedure, the encrypting node maintains an estimate of the tunnel MTU. Network operators can configure the tunnel MTU on the encrypting node. Alternatively, they can configure the encrypting node to automatically discover and maintain a running estimate of the tunnel MTU. Today, when a encrypting node is configured to automatically discover the tunnel MTU, it executes ICMP-based PMTU Discovery (PMTUD) [RFC1191] [RFC8201] procedures. Having discovered the PMTU, it calculates the tunnel MTU by subtracting the IPSec tunnel overhead from the PMTU.

The above-mentioned ICMP-based PMTUD procedures are susceptible to attack [I-D.roca-ipsecme-ptb-pts-attack]. An attacker can forge an ICMP PTB message, setting the MTU to a low value. When the encrypting node receives the forged ICMP PTB message, it decreases its estimate of tunnel MTU, causing unnecessary fragmentation. Therefore, many IPsec implementations do not implement tunnel MTU discovery at all.

This document describes Packetization Layer PMTU Discovery (PLPMTUD) procedures for IPSec tunnels. In these procedures, the encrypting node discovers and maintains a running estimate of the tunnel MTU. In order to do this, the encrypting nodes sends Probe Packets of various size through the IPSec tunnel. If the size of Probe Packet exceeds the tunnel MTU, a downstream node discards the packet and sends an ICMP PTB message to the encrypting node. The encrypting node ignores the ICMP PTB message. If the size of the Probe Packet does not exceed the tunnel MTU and the decrypting node receives the Probe Packet, the decrypting node sends an Acknowledgement Packet to encrypting node through the IPSec tunnel. The Acknowledgement Packet indicates the size of the Probe Packet. Unlike ICMP PTB messages, this Acknowledgement Packet cannot be forged.

The procedures described in this document are applicable to IPSec tunnels that are signaled by Internet Key Exchange version 2 (IKEv2) [RFC7296] and provide authentication services.

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. PLPMTU Discovery Procedures

3.1. Method of Operation

A special loopback interface is configured on the encrypting and decrypting nodes. In this document, these loopback interfaces are called the IPSec PLPMTUD interfaces.

The encrypting node executes IKEv2 procedures to signal an IPSec tunnel between itself and a decrypting node. The IPSec tunnel MUST provide authentication services. It MAY also provide privacy services. If the outermost header of the IPSec tunnel is an IPv4 header, the DF bit must be set. IKEv2 endpoints MUST exchange traffic selectors advertising their IPSec PLPMTUD interface

addresses. Implementations MUST ensure that traffic from one IPSec PLPMTUD address to another traverses the appropriate tunnel using the correct security association.

As part of the tunnel establishment process, the encrypting node produces an initial estimate of the tunnel MTU. The encrypting node's initial estimate of the tunnel MTU is equal to its initial PMTU estimate minus IPSec tunnel overhead, where:

- o The initial PMTU estimate is equal to the MTU of the first link along the path between the encrypting node and the decrypting node.

- o IPSec tunnel overhead is the maximum number of bytes required for padding (by the encryption algorithm) plus the number of bytes required for IPSec encapsulation.

This initial estimate may be greater than the actual tunnel MTU.

Having established the IPSec tunnel, the ingress node begins to refine its estimate of the tunnel MTU. It MAY pass traffic through the tunnel as it refines the tunnel MTU estimate.

In order to refine its estimate of the tunnel MTU, the ingress node executes the Packetization Layer PMTU Discovery (PLPMTUD) procedures described in Section 4 of [I-D.fairhurst-tsvwg-datagram-plpmtud]. When applied to IPSec tunnels, these procedures can be summarized as follows:

- o The encrypting node sends Probe Packets of various size through the IPSec tunnel.

- o If the size of the Probe Packet exceeds the tunnel MTU, a downstream device drops the packet and sends an ICMP Packet Too Big (PTB) message to the encrypting node. The encrypting node ignores the ICMP PTB message.

- o If the Probe Packet reaches the decrypting node, the decrypting node acknowledges receipt of the Probe Packet.

Section 3.2 of this document describes the Probe Packet. Section 3.3 of this document describes how the decrypting node acknowledges receipt of the Probe Packet.

3.2. PLPMTUD Probe

The encrypting node can probe the IPSec tunnel using an IPv4 packet or an IPv6 packet. Figure 1 depicts the IPv4 Probe Packet while Figure 2 depicts the IPv6 Probe Packet. In either case, the encrypting node forwards the Probe Packet through the IPSec tunnel.

```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |Version|  IHL  |Type of Service|          Total Length         |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |         Identification        |Flags|      Fragment Offset    |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |  Time to Live |    Protocol   |         Header Checksum       |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   IPv4 |                       Source Address                          |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |                    Destination Address                        |
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |           Source Port         |         Dest Port             |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |           UDP Length          |        UDP Checksum           |
    UDP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |P|                        Reserved                             |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |                                                               |
     |  //                           Padding                           //
     |  |                                                               |
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: IPv4 Probe Packet

```
         0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |Version| Traffic Class |           Flow Label                  |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |         Payload Length        |  Next Header  |   Hop Limit   |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |                                                               |
     |  +                                                               +
     |  |                                                               |
     |  +                         Source Address                        +
   IPv6 |                                                               |
     |  +                                                               +
     |  |                                                               |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |                                                               |
     |  +                                                               +
     |  |                                                               |
     |  +                      Destination Address                      +
     |  |                                                               |
     |  +                                                               +
     |  |                                                               |
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |           Source Port         |          Dest Port            |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |           UDP Length          |        UDP Checksum           |
    UDP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |P|                       Reserved                              |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |                                                               |
     |  //                           Padding                            //
     |  |                                                               |
      -+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: IPv6 Probe Packet

Regardless of whether the Probe Packet is IPv4 or IPv6:

- o The Source Address is the encrypting node's IPSec PLPMTUD interface address.
- o The Destination Address is the decrypting node's IPSec PLPMTUD interface address.
- o The Source Port is chosen from the dynamic port range (49152-65535) [RFC6335]

o The Destination Port is equal to IPSec PLPMTUD. (Value TBD by IANA).

o The P-bit is set to indicate that this is a Probe Packet.

o The Reserved Field MUST be set to zero and MUST be ignored upon receipt.

o The Padding field is used to vary the size of the packet.

3.3. PLPMTUD Acknowledgement

When the decrypting node receives a Probe Packet, it returns an Acknowledgment Packet. The Acknowledgment Packet can be an IPv4 packet or an IPv6 packet. Figure 3 depicts the IPv4 Acknowledgment Packet while Figure 4 depicts the IPv6 Acknowledgment Packet. In either case, the decrypting node forwards the Acknowledgment Packet through the IPSec tunnel that connects it to the encrypting node.

```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |Version|  IHL  |Type of Service|          Total Length         |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |         Identification        |Flags|      Fragment Offset    |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     | |  Time to Live |    Protocol    |         Header Checksum       |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   IPv4 |                       Source Address                          |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |                    Destination Address                        |
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |           Source Port         |          Dest Port            |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    UDP |           UDP Length          |         UDP Checksum          |
     |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  |P|         Reserved            |       Probe Length            |
      - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: IPv4 Acknowledgement Packet

```
       0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  |Version| Traffic Class |           Flow Label                  |
   |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  |         Payload Length        | Next Header   |   Hop Limit   |
   |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  |                                                               |
   |  +                                                               +
   |  |                                                               |
   |  +                     Source Address                            +
IPv6  |                                                               |
   |  +                                                               +
   |  |                                                               |
   |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  |                                                               |
   |  +                                                               +
   |  |                                                               |
   |  +                  Destination Address                          +
   |  |                                                               |
   |  +                                                               +
   |  |                                                               |
    - +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  |           Source Port         |         Dest Port             |
   |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 UDP  |           UDP Length          |        UDP Checksum           |
   |  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  |P|         Reserved            |       Probe Length            |
    -  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: IPv6 Acknowledgement Packet

Regardless of whether the Acknowledgment Packet is IPv4 or IPv6:

- o The Source Address is copied from the Destination Address of the corresponding Probe Packet

- o The Destination Address is copied from the Source Address of the corresponding Probe Packet

- o The Source Port is copied from the Destination Port of the corresponding Probe Packet

- o The Destination Port copied from the Source Port of the corresponding Probe Packet

- o The P-bit is clear to indicate that this is an Acknowledgement Packet.

- o The Reserved Field MUST be set to zero and MUST be ignored upon receipt.

- o The Probe length represents the total length of the corresponding Probe Packet, measured in bytes and not counting IPSec overhead

4. Security Considerations

The procedures described herein are an improvement upon ICMP-based PMTUD procedures because unlike ICMP PTB messages, the Acknowledgement Packets described herein cannot be forged.

The decrypting node MUST protect the encrypting node from forged Acknowledgement Packets. Therefore, the decrypting MAY originate packets whose source address is its PLPMTUD interface address. However, it MUST NOT forward packets whose source address is its PLPMTUD interface address.

5. ECMP Considerations

Packets traversing a network, with multi paths (ECMP), would end up picking the lowest MTU available in any of the ECMP paths, when the proposed solution is employed (assuming that paths have different MTU values for the sake of analysis). This might cause some additional load on the encryption end, due to the lower MTU level fragmentation. This wouldn't be a major issue, as even otherwise, these loads would have got processed on the receiving side (decryption side) for reassembly and holding the packets in memory. It is worth noting that, at the encryption side it is more of 'stateless' action in terms of packet fragmentation is concerned as compared to at decryption side it is more of a 'stateful' action, where in, it need to maintain the fragments queue for reassembly. Moreover, reassembly node has no control over arrival of the fragments. So, when a choice has to be made for loading the end between encryption and decryption end, it is always better to load the encryption side due to the fact that the operation is stateless and less costly to perform comparatively.

6. IANA Considerations

IANA is request to allocate a UDP port called "IPSec PLPMTUD" from the Registered Port Range (1024 to 49151).

7. Acknowledgements

Thanks to Yoav Nir, Joe Touch, and Dan Wing for their review and comments.

8. References

8.1. Normative References

[I-D.fairhurst-tsvwg-datagram-plpmtud]
Fairhurst, G., Jones, T., Tuexen, M., and I. Ruengeler, "Packetization Layer Path MTU Discovery for Datagram Transports", draft-fairhurst-tsvwg-datagram-plpmtud-02 (work in progress), December 2017.

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, DOI 10.17487/RFC0791, September 1981, <https://www.rfc-editor.org/info/rfc791>.

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, DOI 10.17487/RFC0792, September 1981, <https://www.rfc-editor.org/info/rfc792>.

[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>.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <https://www.rfc-editor.org/info/rfc4301>.

[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>.

[RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", BCP 165, RFC 6335, DOI 10.17487/RFC6335, August 2011, <https://www.rfc-editor.org/info/rfc6335>.

[RFC7296] Kaufman, C., Hoffman, P., Nir, Y., Eronen, P., and T. Kivinen, "Internet Key Exchange Protocol Version 2 (IKEv2)", STD 79, RFC 7296, DOI 10.17487/RFC7296, October 2014, <https://www.rfc-editor.org/info/rfc7296>.

[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>.

8.2. Informative References

[I-D.roca-ipsecme-ptb-pts-attack]
Roca, V. and S. Fall, "Too Big or Too Small? The PTB-PTS ICMP-based Attack against IPsec Gateways", draft-roca-ipsecme-ptb-pts-attack-00 (work in progress), July 2015.

Authors' Addresses

Shibu Piriyath
Juniper Networks
Elnath-Exora Business Park
Bangalore, KA 93117
India

Email: spiriyath@juniper.net

Umesh Mangla
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
USA

Email: umangla@juniper.net

Nagavenkata Suresh Melam
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
USA

Email: nmelam@juniper.net

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

Email: rbonica@juniper.net