IPsecme
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2025

D. Liu, Ed.
D. Migault
R. Liu
C. Zhang
Ericsson

3 March 2025

IKEv2 Link Maximum Atomic Packet and Packet Too Big Notification Extension
draft-liu-ipsecme-ikev2-mtu-dect-09

Abstract

This document defines the IKEv2 Link Maximum Atomic Packet and Packet Too Big Extension to limit reassembly operations being performed by the egress security gateway.

This extension enables an egress security gateway to notify its ingress counterpart that fragmentation is happening or that the received (and potentially reassembled) ESP packet is too big and thus cannot be decrypted. In both cases, the egress node provides Maximum Transmission Unit (MTU) information. Such information enables the ingress node to configure appropriately its Tunnel Maximum Transmission Unit - also designated as MTU or Tunnel MTU (TMTU) - to prevent fragmentation or too big packets to be transmitted.

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 4 September 2025.

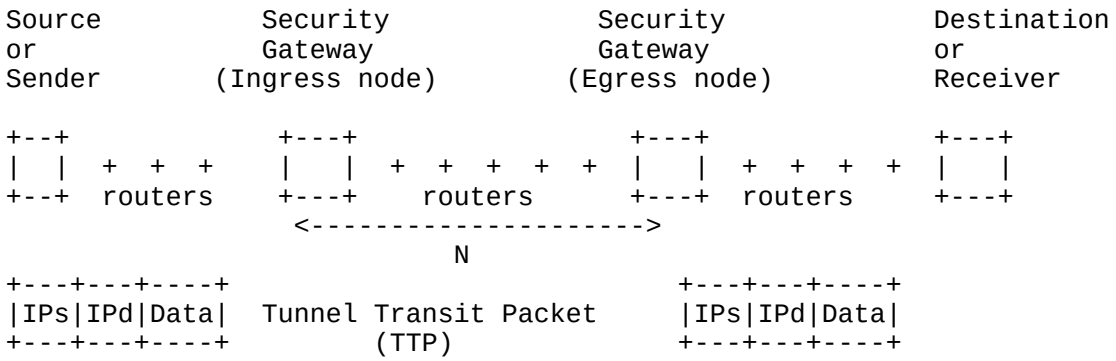
Copyright Notice

Table of Contents

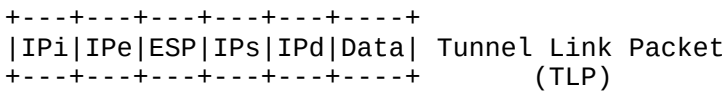
1. Introduction

Reassembling fragments at the egress security gateway requires additional resources which under heavy load results in service degradations. Then, as detailed in [RFC4963], [RFC6864] or [RFC8900], fragmentation is considered fragile and not sufficiently robust at high data rates. Typically, the 16-bit IPv4 identification field is not large enough to prevent duplication making fragmentation not sufficiently robust at high data rates. In IPv6 the 32 byte identification field reduces such collisions.

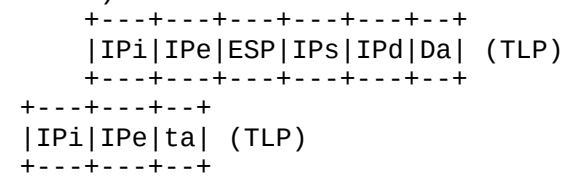
Figure 1 depicts various fragmentation scenarios that can occur when Tunnel Transit Packets (TTP) are encapsulated over an IPsec tunnel. The IPsec packets exchanged between the ingress and the egress security gateway are designated as Tunnel Link Packet (TLP).



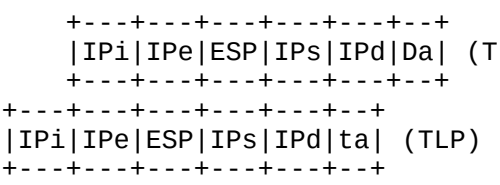
1) No fragmentation



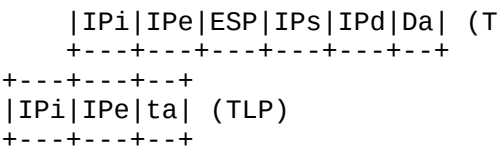
2) Mid-tunnel (performed by a router on N)
(only for IPv4 DF=0 TLP)



3) Inner fragmentation (performed by the Ingress node)
(only for IPv4 DF=0 TTP)



4) Outer fragmentation (performed by the Ingress node)
(IPv4 or IPv6 TLP)



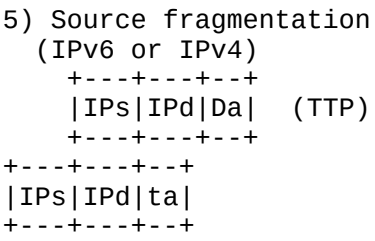


Figure 1: Illustration of Different Type of Fragmentation. IPs (resp. IPi, IPe, IPd) designates the IP addresse of the Source, (resp. the Ingress node, the Egress node, the Destination). The IPsec tunnel is considered as an ESP tunnel. The IP payload is represented as 'Data'. Fragmentation is illustrated in splicing 'Data' into 'Da' and 'ta'. The figure does not show a difference between Data being encrypted or not, and the presence of the ESP header indicates the payload is encrypted.

Reassembling is performed by the egress node in two cases. Firstly when mid tunnel fragmentation happens (see 2) Figure 1) -- in which case the TLP header or outer header is using IPv4 with its Do not Fragment bit set to 0 (DF=0). Secondly when Outer fragmentation is performed by the ingress node (see 4 in Figure 1). The main difference between the two scenarios is that with Outer fragmentation, the ingress node is aware that the egress performs some reassembly operations. Note also that in both cases, reassembling the TLP in itself does not prevent the TTP to be processed by IPsec unless the reassembled TLP exceeds the effective MTU to receive (EMTU_R) - that is the maximum size of the IPsec protected packet that can be accepted by the egress node to perform the ESP encapsulation.

Figure 2 summarizes the various operations performed by the egress node according to the size of an IPsec encapsulated TTP. The optimal size of the TTP is the maximum TTP size that avoids fragmentation and the corresponding TTP size is designated as Tunnel maximum atomic packet (TMAP). The resulting IPsec encapsulated TTP (or LTP) is designated as Link maximum atomic packet (LMAP) LTP. The difference between the two sizes is related to the IPsec encapsulation overhead. LTP smaller than the LMAP will not generate any fragmentation. LTP larger than the LMAP, but smaller than the EMTU_R will be fragmented and reassembled by the egress node. The TTP will be transmitted. LTP (eventually reassembled) with a size greater than EMTU_R cannot be handled by the egress node and will not be transmitted. Note that in this case and unless specified explicitly the link considered is the physical link between the ingress and egress node.

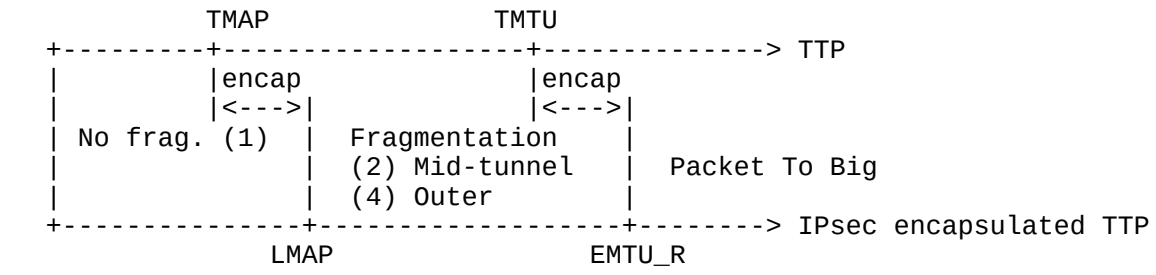


Figure 2: Fragmentation and Packet Too Big as a function of the TTP size or LTP size. encap designates the IPsec overhead.

This document enables a egress node to inform the ingress node that:

* a received packet is fragmented (via an IKE LMAP notification)

* a packet received is too big to be processed by IPsec (via an IKE PTB notification)

As depicted in Figure 3, by supporting this extension, the ingress and egress node commit themselves in optimizing the processing of the IPsec tunnel and in limiting reassembly operation being performed at the egress node. More specifically, the ingress security gateway limits as much as possible the use of outer fragmentation and commits to set their TMTU value so that TTP are not fragmented/reassembled. In addition, for TTP with IPv4 addresses and DF=0, the ingress node commits to perform inner fragmentation to prevent reassembly at the egress node.

The mechanism is especially useful when the tunnel between the ingress and egress nodes is using IPv4 outer IP addresses with DF=0 as the fragmentation may occur while the ingress may not be aware of it.

The mechanisms enables also enable to inform the IPsec encapsulated TTP is too large and cannot be processed by IPsec. A LTP packet may be to big if it exceeds the LMTU or the EMTU_R. When one of these boundaries is exceeded, both are returned to the ingress node so, the ingress node can clearly understand why the packet has been dropped.

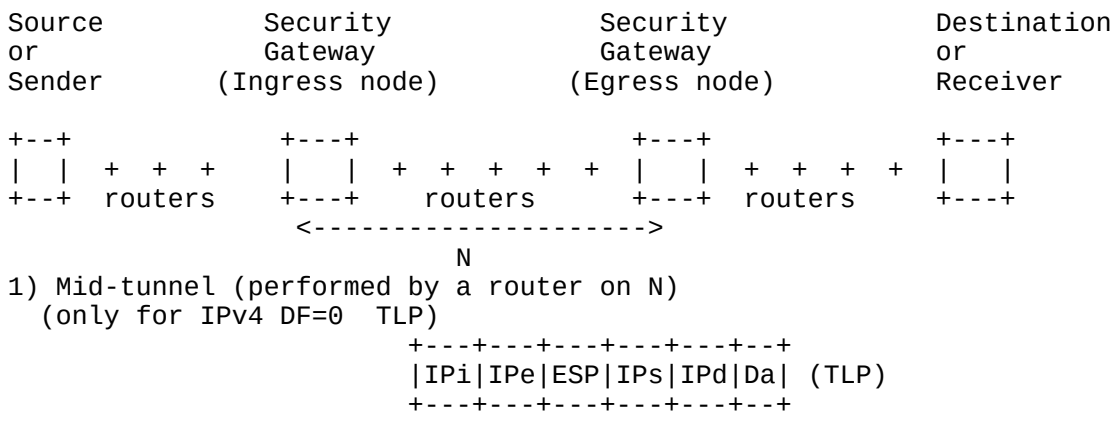
A packet may meet the LMTU requirement, but may not meet the EMTU_R requirement.
With IPv4 outer IP addresses with DF=0, the egress Security Gateway may receive a serie of fragments that individually do not exceed the MTU (and as such do not generate any LTP ICMP PTB) reassemble a fragmented IPsec encapsulated. That reassembled packet may exceed the size to be processed by IPsec. Similarly, even without

fragmentation, the LTP MTU may exceed the capacity of the IPsec hardware in which case a IPsec encapsulated may not generate a LTP ICMP PTB but may not be further processed. Even though MTU values may be sent via a LTP ICMP PTB message, we do envision that the communication of the LMTU via an IKE notification may provide some advantages over simply relying on LTP ICMP messages. However, please note these advantages are only secondary and only mentioned here as an information. At first the IKE channel is authenticated channel that ensures the LMTU is 1) received by the ingress Security Gateway, 2) is integrity protected and 3) can be appropriately bound to the corresponding SA. An ICMP PTB message when sent from the egress Security Gateway is not protected and when UDP encapsulation is used (see {sec-sec}) and may not carry the SPI of the IPsec encapsulated TTP (see {sec-df1}). However, these advantages are only provided by the egress interface and so that mechanism cannot be used by any other node, and so cannot be used for PMTU discovery. Secondly, the IKE notification MAY be sent together with the TTP ICMP PTB sent by the router. This may improve the network latency and optimizes the use of the security gateway resources, especially when the SA has been set for multiple Sources. In fact the TTP ICMP PTB sent by the egress router is sent specifically to the sending source, and so one can expect the ICMP PTB being sent to every source. By sending the notification via IKE, the information is received by the ingress Security Gateway which can take action globally for all Sources associated to the SA. This prevents the large packet greater than the MTU to be encrypted and sent to the egress router for nothing.

This document does not discuss these optimizations.

1.1. Illustrative example

This section describes an illustrative example to provide a high level overview.



```
                    +---+---+--+
                    |IPi|IPe|ta| (TLP)
                    +---+---+--+
```

2) Egress node detects fragmentation
- a) it collects IPVersion the IP version of the first fragment as well as FragLen, the fragment length
- b1) If all segments can be reassembled and the reassembled packet is properly decrypted a Link Maximum Atomic Packet Notification (LMAP) is sent on the IKEv2 channel.

```
 [IKEv2]
 <---  N( LMAP [ IPVersion, FragLen] )
```

- b2) If the packet is too big and cannot be fully processed, an additional Packet Too Big Notification (PTB) that specifies the Link MTU (LMTU) of the router component of the egress node (on network N) as well as the effective MTU to receive (EMTU_R). Both are configuration parameters.
  An ICMP PTB message may also be sent by the egress node.
  This is considered as independent from the fact that the egress Security Gateway is sending an IKE PTB or not.
  Note that this ICMP may not contain even the SPI, and so is likely to not carry sufficient information to the ingress node, so any action be taken.

```
 [IKEv2]
 <---  N( LMAP [ IPVersion, FragLen] )
       N( PTB [LMTU, EMTU_R]
 [ESP]
 <--- ESP( ICMP PTB )
```

3) Upon receiving the LMAP or optionally the ingress node
  a) Update the TMTU so that the Source performs source fragmentation with TTP packets that are not fragmented.

  Source fragmentation
  (IPv6 or IPv4)

```
    +---+---+--+
    |IPs|IPd|Da|  (TTP)
    +---+---+--+
+---+---+--+
|IPs|IPd|ta|
+---+---+--+
```

  b) Performs inner fragmentation TTP packets that exceeds the TMTU and will generate some fragments.

  Inner fragmentation (performed by the Ingress node)
  (only for IPv4 DF=0 TTP)

```
               +---+---+---+---+---+--+
```

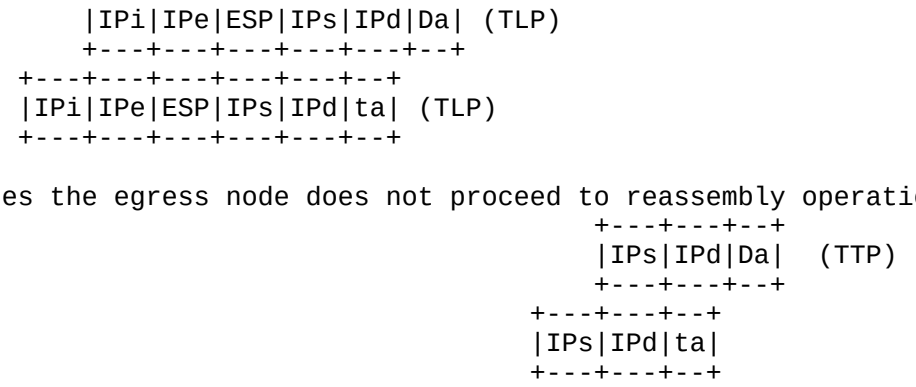


Figure 3: Overview mechanism

1.2. Related works

This work differs from [I-D.ietf-intarea-tunnels] in that [I-D.ietf-intarea-tunnels] which considers the router component - carrying the TTP - and the interface component - handling LTP - independent. Independence of the Tunnel MTU (for TTP) and link layer MTU for (LTP) is provided by performing outer fragmentation when needed. [RFC4301] takes another view considering the router component can adapt to the specific needs of the interface component. In our case, the router MTU can be changed so the source sends PTP of an expected TMAP size. Note however, that a significant difference between MPA and MTU is that fragmentation is considered as a normal operation and that ICMP Packet Too Big (PTB) is only expected when the router is not able to handle the packet - that is when the (reassembled) packet exceeds the MTU (or more explicitly the effective MTU to receive (EMTU_R) or the Link MTU (LMTU)).

The extension described in this document follows [RFC8900]' s recommendations where each layer handles fragmentation at their layer and to reduce the reliance on IP fragmentation to the greatest degree possible. This document does not describe a Path MTU Discovery (PMTUD) procedure [RFC1191] nor an Execute Packetization Layer PMTUD (PLMTUD) [RFC4821] procedure. PLMTUD work has been started in [I-D.spiriyath-ipsecme-dynamic-ipsec-pmtu]. This document remains focused on providing information on the state of the egress node to the ingress node. This Information is fragmentation related and includes the notification that fragmentation is being observed as well as the EMTU_R and LMTU value - which are configuration parameters. This document lets the ingress node interpret these pieces of information and take the necessary actions. PLMTUD is much more complex especially as IPsec considers multiple channels such as IKE and IPsec protected data and is required to handle black holing scenarios.

Sending LMTU when a too big LTP is received by the egress router is similar to an ICMP PTB except that the message is carried over the IKEv2 channel, and we ensure that sufficient information (like the SPI) is provided to the ingress node so the appropriated traffic selectors can be identified by the ingress node - see Section 1.3 for more details. EMTU_R is not sent by a LTP ICMP PTB packet.

1.3. Why not using DF=1 to avoid Mid fragmentation ?

One can reasonably question why setting the IPv4 DF=1 is not sufficient to avoid fragmentation. While DF=1 avoids fragmentation, it can easily fall into black holing scenarios, unless one can ensure that ICMP PTB messages are effectively received with sufficient information by the ingress node to take appropriate actions. This is not the case in practice, which is the reason DF is very commonly set to 0.

Suppose the Don’t Fragment bit to 1 in the IPv4 Header of the Tunnel Link Packet. If that packet becomes larger than the link Maximum Transmission Unit (LMTU), the packet is dropped by an on-path router and an ICMPv4 message Packet Too Big (PTB) [RFC0792] is returned to the sending address. The ICMPv4 PTB message is a Destination Unreachable message with Code equal to 4 and was augmented by [RFC1191] to indicate the acceptable MTU. Unfortunately, one cannot rely on such a procedure as in practice some routers do not check the MTU and as such do not send ICMPv4 messages. In addition, when ICMPv4 messages are sent these messages are unprotected, and may be blocked by firewalls or ignored. This results in IPv4 packets being dropped without the security gateways being aware of it which is also designated as black holing. To prevent this situation, IPv4 packets often set their DF bit set to 0. In this case, as described in [RFC0792], when a packet size exceeds its MTU, the node fragments the incoming packet in multiple fragments.

In addition to the above reasons DF=1 is not appropriate for ESP, there is another important reason that ICMP does not work almost completely for ESP.

Figure 4 describes the ICMPv4 PTB as defined in [RFC1191] and to be useful to the ingress node, the ICMP PTB should at least carry the SPI that would identify the incoming traffic associated to that ICMP PTB. This information is carried in the “Internet Header + 64 bits of Original Datagram Data” field which in our case carries the tunnel IP header with an additional 8 bytes.

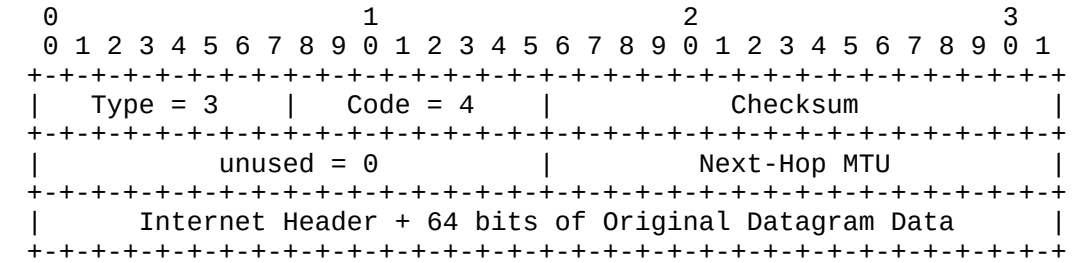


Figure 4: ICMP PTB

When ESP is encapsulated in UDP or TCP, the 8 bytes correspond to the UDP header and the TCP header is even larger. As a result, the SPI cannot be derived from the 8 bytes being provided and the ingress node cannot identify the traffic selector and proceed to the next step.

Note also that when the SPI is identified, the ingress node may not exactly know which Source has generated the ICMP PTB as the SA traffic selector is defined by a range of IP addresses and as such may contain multiple Sources. The Path MTU is propagated to the Source as described in [RFC4301], Section 8.2, that is to say by by updating the PMTU information of the SA and upon receiving a packet that matches that SA and whose size exceeds the PMTU of the SA, discard that packet and sends back an ICMP PTB message to the Source.

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.

Tunnel, Ingress node, Egress node, Ingress Interface, Egress interface, Tunnel Transit Packets (TTP), Tunnel Link Packet (TLP), Link MTU (LMTU), Tunnel MTU (TMTU), Tunnel maximum atomic packet (TMAP), effective MTU to receive (EMTU_R) are defined in [I-D.ietf-intarea-tunnels].

3. Link Maximum Atomic Packet and Packet Too Big Support Negotiation

During an IKEv2 negotiation, the initiator and the responder indicate their support for the Link Maximum Atomic Packet and Packet Too Big Extension by exchanging the LMAP_AND_PTB_SUPPORTED notifications. This notification MUST be sent in the IKE_AUTH exchange (in case of multiple IKE_AUTH exchanges - in the first IKE_AUTH message from initiator and in the last IKE_AUTH message from responder). If both the initiator and the responder send this notification during the IKE_AUTH exchange, peers may notify each other with an IPv4 Link Maximum Atomic Packet Notification when fragmentation is observed. Upon receiving such notifications, the peers may take the necessary actions to prevent such fragmentation to occur.

```
Initiator                          Responder
-------------------------------------------------------------------
HDR, SA, KEi, Ni -->
                          <-- HDR, SA, KEr, Nr
HDR, SK {IDi, AUTH,
     SA, TSi, TSr,
     N(LMAP_AND_PTB_SUPPORTED)} -->
                          <-- HDR, SK {IDr, AUTH,
                              SA, TSi, TSr,
                              N(LMAP_AND_PTB_SUPPORTED)}
```

4. Sending a Link Maximum Atomic Packet Notification

The egress security gateway detects fragmentation occurred when it receives an initial fragment; e.g. with the Flags' More Fragment Bit set to 1 and the Fragment Offset set to 0. Upon receiving such a packet, the egress node determines the IP version (IPVersion) and the fragment length FragLen). For an IPv4 packet, FragLen is the Total Length field (see [RFC0791]). For an IPv6 packet FragLen is the Payload Length (see [RFC8200], Section 3. Note that these values have different meanings as with an IPv6 fragment, FragLen does not include the IPv6 header but only the payload.

The egress node sends the LMAP notification payload that contains IPVersion and FragLen.

The egress node SHOULD send a maximum of one LMAP notification per (reassembled) received packet. However, since this extension is especially expected on nodes dealing with high traffic rates, the notification is expected to be sent at reasonable rates per Security Associations. More specifically, the use of the IKEv2 provides a reliable channel which makes sending redundant notifications unnecessary. Then, the notification rate needs to account for the time the egress node adjusts the TMTU, and that TMTU remains implemented. More details are provided in Section 10.

```
Egress Security Gateway                 Ingress Security Gateway
-----------------------------------------------------------------
HDR SK { N(LMAP)} -->
```

5. Receiving a Link Maximum Atomic Packet Notification

Upon receiving a LMAP notification, the ingress node derives the tunnel MAP (TMAP) from the Link MAP (LMAP) derived by the FragLen and IPVersion.

```
if IPVersion is 4:
  LMAP = FragLen
elif IPVersion is 6:
  LMAP = FragLen + 40
TMAP = LMAP - outer IP header - encapsulation overhead
```

Figure 5: Computation of TMAP

where

IPVersion: The IP version of the fragment provided by the LMAP notification (see Section 4).

FragLen: The Fragment length provided by the LMAP notification (see Section 4).

LMAP: For an IPv4 packet, LMAP is directly provided by the fragment length of the LMAP Notification. For an IPv6 packet, LMAP needs to adds the IPv6 Header length (40 bytes) to the fragment length of the LMAP Notification.

outer IP header: The IP header of the LTP encapsulation overhead: contains the ESP header, the ESP Trailer including the variable Pad field. When the padding is minimizing the Pad Len, the encapsulation header is set to 14 (+ the size of the ICV). The overhead SHOULD also estimate IP options or IP extensions.

The ingress security gateway SHOULD propagate the TMAP as the tunnel MTU back to the Source so the size of future TTP packets does not exceed the TMAP - eventually performing source fragmentation. To do so, the ingress node sets the LMTU to TMAP for all traffic designated by the SA. In this case the LMTU is the MTU associated with the link of the router interface of the ingress node that faces the Source's network. Upon receiving a TLP larger than the TMAP, the packet is discarded and an ICMP PTB message is returned to the Source which then performs Source Fragmentation (5) (See Section 8.2.1. of [RFC4301]). It is worth mentioning that only future packets will be impacted, and not those causing fragmentation.

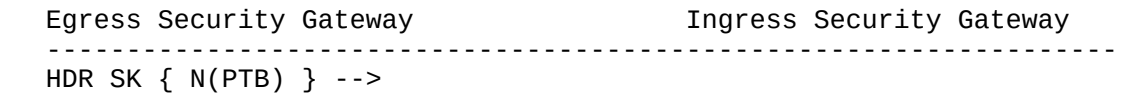
When the TLP is an IPv4 packet with DF=0, the ingress node SHOULD perform Source Fragmentation of the TTP, also represented as Inner Fragmentation (3), sending chunks that do not exceed TMAP.

Figure 11 in Section 4.2.2 of [I-D.ietf-intarea-tunnels] with tunnel MTU set to TMAP achieves both recommendations, while Figure 12 in Section 4.2.2 describes the inner fragmentation.

6. Sending a Packet Too Big Notification

A packet can be rejected because the size of the LTP exceeds the LMTU (of the router component) or when the (reassembled) LTP exceeds the EMTU_R (of the interface component) and so IPsec decapsulation cannot be done.

When the LTP size exceeds the EMTU_R, the egress node SHOULD send a Packet Too Big (PTB) notification that includes the EMTU_R and the LMTU. In addition, if the packet results from a reassembly operation, the egress node MUST send a LMAP notification with the LMAP. If the packet does not result from a reassembly operation, the egress node MUST NOT send a LMAP notification.



7. Receiving a Packet Too Big Notification

Upon receiving a PTB notification, the egress node computes the Tunnel MTU (TMTU) as follows:

```
TMTU = EMTU_R - outer IP header - encapsulation overhead
if LMAP notification is not received:
  TMAP is derived from the LMAP notification
else:
  TMAP = LMTU - outer IP header - encapsulation overhead
TMAP = min( TMAP, TMTU )
```

Figure 6: Computation of TMTU

with the same notation as in Figure 5:

EMTU_R: The value provided in the PTB notification related to the MTU associated to the egress interface (see Section 6)

LMTU : The value provided in the PTB notification related to the LMTU associated to the egress router (see Section 6)

The ingress node SHOULD proceed with TMAP as described in Section 5.

The ingress node MUST ensure the size of the TTP do not exceed the computed TMTU and MUST ensure the size of the LTP do not exceed the LMTU provided in the PTB notification.

8. Payload Description

Figure 7 illustrates the Notify Payload packet format as described in Section 3.10 of [RFC7296] with a 4 bytes path allowed MTU value as notification data. This format is used for both the LMAP_AND_PTB_SUPPORTED, LMAP and PTB notifications.

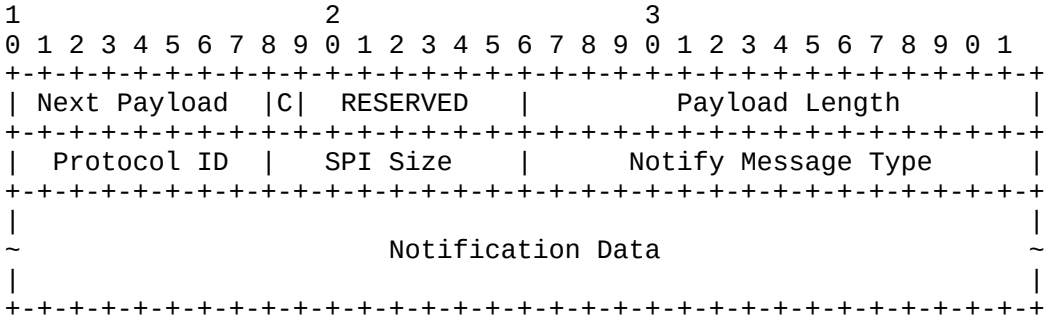


Figure 7: Notify Message Format

The fields Next Payload, Critical Bit, RESERVED and Payload Length are defined in [RFC7296]. Specific fields defined in this document are:

Protocol ID (1 octet): set to zero.

SPI Size (1 octet): set to zero.

Notify Message Type (2 octets): Specifies the type of notification message. It is set to TBD1 for the LMAP_AND_PTB_SUPPORTED notification, TBD2 for the LMAP notification and TBD3 for the PTB notification.

Notification Data: Specifies the data associated to the notification message. It is empty for the LMAP_AND_PTB_SUPPORTED notification or a 4 octets that contains the MTU value for the LMAP and PTB notification - as represented in Figure 8 and Figure 9.

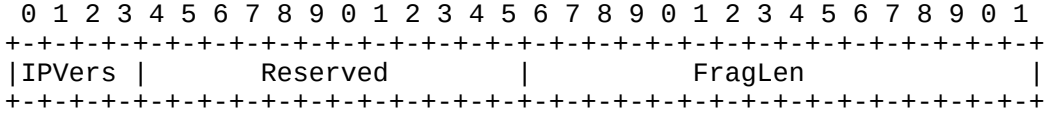


Figure 8: Notification Data for LMAP

with:

IPVersion (4 bits): The IPversion of the received packet

Reserved: Reserved bytes MUST be set by the egress node to zero and MUST be ignored by the ingress node.

FragLen (2 bytes): The FragLen value (see Section 4)

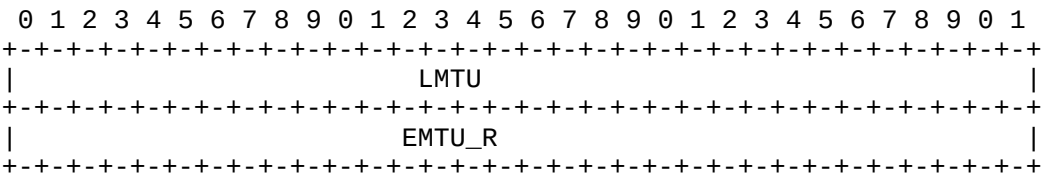


Figure 9: Notification Data for PTB

with:

LMTU (4 bytes): The LMTU of the egress router

EMTU_R (4 bytes): The EMTU_R of the egress interface

9. IANA Considerations

IANA is requested to allocate two values in the "IKEv2 Notify Message Types - Status Types" registry (available at https://www.iana.org/assignments/ikev2-parameters/ikev2-parameters.xhtml#ikev2-parameters-16) with the following definition:

Value	NOTIFY MESSAGES - STATUS TYPES
TBD1	LMAP_AND_SUPPORTED
TBD2	LMAP
TBD3	PTB

10. Security Considerations

This document defines an IKEv2 extension to enable an egress node to notify an ingress node that fragmentation is happening as well as the observed fragment length. In addition, the extension also enables an egress node to notify an ingress node that a packet too big has been discarded, together with some complementary information to appropriately update the MTU.

These pieces of information are transferred over the authenticated IKEv2 channel which ensures the origin of the message. Assuming the egress node is trusted, the ingress node can trust what is being reported effectively observed (like fragmentation is happening, the observed fragment length, a packet too big has been received) by the egress node and that some information are effectively accurate such as the egress LMTU and EMTU_R. When fragmentation happens and a LMAP notification is being sent, the egress node MUST send the

notification once the reassembled packet has been decapsulated. This ensures that fragmentation has been performed over an authenticated TLP and ensures the TLP has not been forged by any attacker. With IPv6, only outer fragmentation is permitted so, the ingress node can validate the provided information. However, sending the notification after the IPsec decapsulation enables the egress node to detect potential injection attacks and prevent sending an unnecessary notification, that may be part of a DDoS attack targeting the ingress node itself. With IPv4 an attacker could set the DF=0 which would allow any mid tunnel fragmentation. IPsec (ESP or AH) does not cover the DF flag, so the egress cannot trust the fragment length observed has not been forged, and the security considerations related to MTU discovery [RFC0791], [RFC8900], [RFC4963], [RFC6864], [RFC1191] apply here. Note that information carried by the LMAP notification is never carried by ICMP, and all LMAP may share with ICMP is that this information will be used to update the MTU.

The egress node may not be able to decrypt the encrypted TTP packet if the full encrypted TTP cannot be built. One possibility is that too many fragments are being sent over a too long period of time (slowloris-like attacks) (see [RFC8900], Section 3.7). Another possibility is that one fragment exceeds the LMTU or that the reassembled (unverified) encrypted TTP exceeds the EMTU_R. In both cases, a PTB notification SHOULD be sent and if fragmentation is observed a LMAP MUST be sent together with the PTB notification. Information carried by the PTB (LMTU and EMTU_R) can be trusted. Without this extension this information would have been carried by ICMP. In many deployments, the ICMP channel may be unprotected and ICMP packets may be discarded by firewalls and never reach the egress node. In addition, the description provided by [I-D.ietf-intarea-tunnels] tends to indicate that the ICMP channel remains between the router components of the ingress and egress nodes and as such are not provided to the interfaces component. Finally, as detailed in Section 1.3 an ICMP PTB message contains a portion of the encrypted ESP packet, which may not sufficient to deduce the SPI and associated traffic selectors, and as such prevent the ingress node to identify the traffic flow that generates the fragmentation. In any case, this results in the information not being available to take the appropriate action. Sending the PTB notification over ICMP solves these issues and eases the correlation with the LMAP notification. In terms of trust, when sufficient information may be sent both on the IKEv2 channel and via a protected ICMP PTB message, the use of the PTB notification achieves similar trust as the one observed with an ICMP PTB message sent over an IPsec protected channel. For that reason, the ICMP messages SHOULD be protected by IPsec. The use of two different paths may provide some additional reliability as the same information is taking two different paths and that IKEv2 windows ensures the the information is received - as

opposed to the (encrypted) ICMP message that can be dropped. However, information carried by the LMAP notification cannot be trusted and similar security considerations related to MTU discovery [RFC0791], [RFC8900], [RFC4963], [RFC6864], [RFC1191] apply here.

During high packet rates, this notification for each of these packets is likely to be used by an attacker to trigger a DDoS attack to both egress and ingress nodes. As a result, the egress node SHOULD be able to configure the maximum rate at which the notifications are sent. This includes the ability to indicate that LMAP notifications (without PTB) are not sent when the outer IP addresses are of version IPv6. The reasoning is that with IPv6, the egress node observes outer fragmentation, in which case the ingress node is already aware of it. In addition, an egress node SHOULD be able to configure a threshold for number of alerts per SAs before a notification is sent, a rate limit per SA.

11. Acknowledgements

The authors would like to thank Magnus Westerlund, Paul Wouters, Joe Touch, Tero Kivinen for his reviews and valuable comments and suggestions.

12. References

12.1. Normative References

[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>.

[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>.

[RFC6864] Touch, J., "Updated Specification of the IPv4 ID Field", RFC 6864, DOI 10.17487/RFC6864, February 2013, <https://www.rfc-editor.org/info/rfc6864>.

[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>.

[RFC8900] Bonica, R., Baker, F., Huston, G., Hinden, R., Troan, O., and F. Gont, "IP Fragmentation Considered Fragile", BCP 230, RFC 8900, DOI 10.17487/RFC8900, September 2020, <https://www.rfc-editor.org/info/rfc8900>.

12.2. Informative References

[I-D.ietf-intarea-tunnels]
Touch, J. D. and M. Townsley, "IP Tunnels in the Internet Architecture", Work in Progress, Internet-Draft, draft-ietf-intarea-tunnels-14, 3 November 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-intarea-tunnels-14>.

[I-D.spiriyath-ipsecme-dynamic-ipsec-pmtu]
Piriyath, S., Mangla, U., Melam, N., and R. Bonica, "Packetization Layer Path Maximum Transmission Unit Discovery (PLPMTUD) For IPsec Tunnels", Work in Progress, Internet-Draft, draft-spiriyath-ipsecme-dynamic-ipsec-pmtu-01, 1 March 2018, <https://datatracker.ietf.org/doc/html/draft-spiriyath-ipsecme-dynamic-ipsec-pmtu-01>.

[RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, DOI 10.17487/RFC4963, July 2007, <https://www.rfc-editor.org/info/rfc4963>.

Authors' Addresses

Daiying Liu (editor)
Ericsson
Email: harold.liu@ericsson.com

Daniel Migault
Ericsson
Email: daniel.migault@ericsson.com

Renwang Liu
Ericsson
Email: renwang.liu@ericsson.com

Congjie Zhang
Ericsson
Email: congjie.zhang@ericsson.com