Lossless Path MTU Discovery (PMTUD)
draft-bonica-intarea-lossless-pmtud-00

Abstract
This document describes alternative IPv4 PMTUD procedures that do not prevent IP fragmentation and do no rely on the network’s ability to deliver ICMP Destination Unreachable messages to the source node. This document also defines a new ICMP message. IPv4 nodes emit this new message when they reassemble a fragmented packet.

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 May 1, 2020.

Copyright Notice
Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents.
1. Introduction

For reasons described in [RFC1191], IPv4 source nodes estimate the Path MTU (PMTU) between themselves and destination nodes. An extremely conservative source node estimates the PMTU for each path to be equal to the IPv4 Minimum Link MTU (See Note 1). While such conservative estimates are guaranteed to be less than or equal to the actual PMTU, they are likely to be much less than the actual PMTU. This may adversely affect upper-layer protocol performance.

By executing PMTU Discovery (PMTUD) [RFC1191] procedures, IPv4 source nodes can maintain less conservative PMTU estimates. In PMTUD, 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 non-fragmentable packets to the destination node (see NOTE 2). 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 drops the packet and sends an Internet Control Message Protocol (ICMP) [RFC0792] Destination Unreachable message to the source node. The Code field in the ICMP message is set to (4) "fragmentation needed". The ICMP message also 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 produces a running estimate of the PMTU between a source node and a destination node. Because PMTU is dynamic, the PMTU estimate can be larger than the actual PMTU. In order to detect PMTU increases, PMTUD occasionally resets the PMTU estimate to its initial value and repeats the procedure described above.

Ideally, PMTUD operates as described above. However, PMTUD relies on the network’s ability to deliver ICMP Destination Unreachable messages to the source node. If the network cannot deliver ICMP Destination Unreachable messages to the source node, PMTUD fails and connectivity may be lost.

This document describes alternative PMTUD procedures that do no rely on the network’s ability to deliver ICMP Destination Unreachable messages to the source node. In 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 fragmentable 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, in one piece, through the next link along the path. Therefore, the downstream router fragments the packet and forwards each fragment to the destination node. The destination node reassembles the packet and sends an informational ICMP message to the source node. The informational message indicates that a packet has been reassembled. It also indicates the size of the largest fragment received and contains as much of the original packet as possible without causing the ICMP message to exceed its maximum allowable size (i.e., 576 bytes).

The source node should use information contained by the message to refine its PMTU estimate. Having refined its PMTU estimate, the source node should refrain from sending packet long enough to require fragmentation. However, the message may be lost by the network or ignored by the source node. In this case, the source node may continue to send packets that require fragmentation and reassembly.

In order to detect PMTU increases, the above-mentioned PMTUD procedures occasionally resets the PMTU estimate to its initial value and repeat the procedure described above.

This document defines the new ICMP message, mentioned above. The PMTUD procedures described herein are applicable to IPv4 only, because [RFC8200] does not allow fragmentation by transit nodes.
This document does not update [RFC1191]. A source node can execute the PMTUD procedures described herein in addition to [RFC1191] procedures or instead of [RFC1191] procedures.

NOTE 1: In IPv4, every host must be capable of receiving a packet whose length is equal to 576 bytes. However, the IPv4 minimum link MTU is not 576. Section 3.2 of [RFC0791] states that the IPv4 minimum link MTU is 68 bytes. But for practical purposes, many network operators consider the IPv4 minimum link MTU to be 576 bytes. So, for the purposes of this document, we assume that the IPv4 minimum link MTU is 576 bytes.

NOTE 2: The DF-bit in the IPv4 header distinguishes fragmentable IPv4 packets from non-fragmentable IPv4 packets. If the DF-bit is equal to 0, the packet is fragmentable. If the DF-bit equals 1, the packet is not fragmentable.

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. The ICMP Packet Reassembly Message

IPv4 nodes can emit an ICMP Packet Reassembly message when they reassemble a packet. Figure 1 depicts the ICMP Packet Reassembly message.

```
+-------------------------------+-------------------------------+-------------------------------+-------------------------------+
| Type | Code | Checksum                          |
| Unused | Length | Largest Fragment |
| Original Datagram
```

Figure 1: The ICMP Packet Reassembly Message
Type (8 bits) – Packet Reassembly. Value TBD by IANA.

Code (8-bits) – No Error (0), or Reassembly Error (1).

Checksum (16 bits) – See [RFC0792].

Unused (8 bits) – SHOULD be set to zero by sender. MUST be ignored by receiver.

Length (8 bits) – Length of the padded "original datagram" field, measured in 32-bit words.

Largest Fragment (16-bits) – Size of the largest fragment received, measured in bytes.

Original Datagram (variable length) – As much of the original packet as possible, without exceeding the maximum size of an ICMP message (576 bytes). Must be padded to 32-bit boundary. If Code equals Reassembly Error, this field contains the first fragment.

As per [RFC1812], all ICMP messages, including the ICMP Packet Reassembly message, SHOULD be rate limited.

The Code field is included for informational purposes only. The receiving node SHOULD refine its PMTU estimate, regardless of the value contained by the code field.

4. Security Considerations

Security considerations for the procedures described herein are identical to those described for PMTUD. See Section 8 of [RFC1191]. [RFC5927] offers mitigations.

5. IANA Considerations

IANA is requested to add the following to the "ICMP Type Numbers" registry:

o Type – Packet Reassembled. (Value TBD)

IANA is requested to add the following to the "Type TBD – Packet Reassembled" subregistry:

o 0 – No error

o 1 – Reassembly error
6. Acknowledgements

Thanks to TBD for their careful review of this document.

7. References

7.1. Normative References


7.2. Informative References


Authors’ Addresses

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

Email: rbonica@juniper.net

Manoj Nayak
Juniper Networks
Bangalore, KA  560103
India

Email: manojnayak@juniper.net

Bradely Newton
Harvey Mudd College
340 Foothill Blvd.
Claremont, California  91711
USA

Email: bnewton@hmc.edu

Hakan Alpan
Harvey Mudd College
340 Foothill Blvd.
Claremont, California  91711
USA

Email: halpan@hnc.edu

Radon Rosborough
Harvey Mudd College
340 Foothill Blvd.
Claremont, California  91711
USA

Email: rrosborough@hmc.edu
Miles President  
Harvey Mudd College  
340 Foothill Blvd.  
Claremont, California  91711  
USA

Email: mpresident@hmc.edu