Workgroup: BIER Working Group
Internet-Draft:
draft-ietf-bier-path-mtu-discovery-16
Published: 5 November 2023
Intended Status: Standards Track
Expires: 8 May 2024
Authors: G. Mirsky Ericsson
T. Przygienda Juniper Networks
A. Dolganow Nokia

Path Maximum Transmission Unit Discovery (PMTUD) for Bit Index Explicit Replication (BIER) Layer

Abstract

This document describes Path Maximum Transmission Unit Discovery (PMTUD) in Bit Indexed Explicit Replication (BIER) layer.

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 8 May 2024.

Copyright Notice

Table of Contents

1. Introduction

In packet switched networks, when a host seeks to transmit data to a target destination, the data is transmitted as a set of packets. In many cases, it is more efficient to use the largest size packets that are less than or equal to the smallest Maximum Transmission Unit (MTU) for any forwarding device along the routed path to the IP destination for these packets. Such "least MTU" is known as Path MTU (PMTU). Fragmentation or packet drop, silent or not, may occur on hops along the path where an MTU is smaller than the size of the datagram. To avoid any of the listed above behaviors, the packet source must find the value of the least MTU, i.e., PMTU, that will be encountered along the path that a set of packets will follow to reach the given set of destinations. Such MTU determination along a specific path is referred to as path MTU discovery (PMTUD).

[RFC8279] introduces and explains Bit Index Explicit Replication (BIER) architecture and how it supports the forwarding of multicast data packets. [I-D.ietf-bier-ping] introduced BIER Ping as a transport-independent OAM mechanism to detect and localize failures in the BIER data plane. This document specifies how BIER Ping can be used to perform efficient PMTUD in the BIER domain.

1.1. Conventions used in this document

1.1.1. Terminology

This document uses terminology defined in [RFC8279]. Familiarity with this specification and the terminology used is expected.

1.1.2. 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.

2. Problem Statement

[I-D.ietf-bier-oam-requirements] sets forth the requirement to define PMTUD protocol for BIER domain. This document describes the extension to [I-D.ietf-bier-ping] for use in the BIER PMTUD solution.

Current PMTUD mechanisms ([RFC1191], [RFC8201], and [RFC4821]) are primarily targeted to work on point-to-point, i.e. unicast paths. These mechanisms use packet fragmentation control by disabling fragmentation of the probe packet. As a result, a transit node that cannot forward a probe packet that is bigger than its link MTU sends to the packet's source an error notification, otherwise the packet destination may respond with a positive acknowledgment. Thus, possibly through a series of iterations, varying the size of the probe packet, the packet source discovers the PMTU of the particular path.

Applying such existing PMTUD solutions are inefficient for point-to-multipoint paths constructed for multicast traffic. Probe packets must be flooded through the whole set of multicast distribution paths repeatedly until the very last egress responds with a positive acknowledgment. Consider the multicast network presented in Figure 1, where MTU on all links but one (B, D) is the same. If MTU on the link (B, D) is smaller than the MTU on the other links, using existing PMTUD mechanism probes will unnecessarily flood to leaf nodes E, F, and G for the second and consecutive times and positive responses will be generated and received by root A repeatedly.

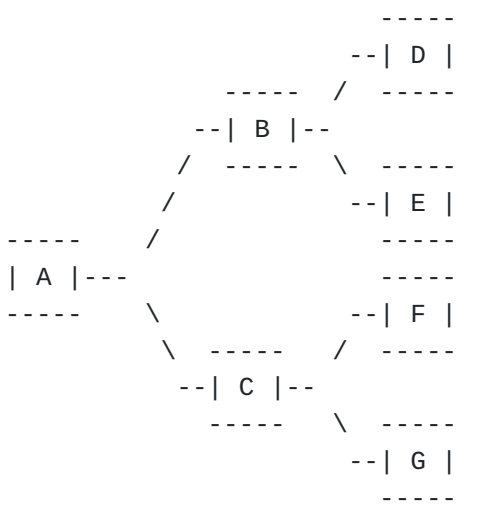

Figure 1: Multicast network

3. PMTUD Mechanism for BIER

A multicast distribution tree connects a BFIR with a set of BFERs via procedures explained in [RFC8279]. The BFIR determines the MTU of this multicast distribution tree by transmitting a series of probe packets from BFIR to the set of BFERs. In the case of ECMP, BFIR MAY test each path by variating the value in the Entropy field. The critical step in the process of Path MTU discovery is the notification of BFIR about the failure at an intermediate BFR to forward the probe packet toward the subset of targeted downstream BFERs. That is achieved by BFR responding with a partial (compared to the one it received in the request) bitmask towards the originating BFIR in error notification. That allows for the retransmission of the next probe with a smaller MTU addressed only toward a smaller set of BFERs downstream from the failed BFR instead of all BFERs within the multicast distribution tree. In the scenario discussed in Section 2, the second and all following (if needed) probes will be sent only to node D because the smaller link MTU of interface B-D. Since the MTU discovery of E, F, and G has been completed already by the first probe, the second, and any of the following probes will not be forwarded to these leaves.

Consider the network displayed in Figure 1 to be a presentation of a BIER domain and all nodes to be BFRs. To discover MTU over BIER domain to BFERs D, F, E, and G BFIR A will use BIER Ping with Data TLV, defined in Section 3.1. Size of the first probe set to M_max determined as minimal MTU value of BFIR's links to BIER domain. As has been assumed in Section 2, MTUs of all links but the link (B, D) are the same. Thus BFERs E, F, and G would receive BIER Echo Request and will send their respective replies to BFIR A. BFR B may pass the

packet which is too large to forward over egress link (B, D) to the appropriate network layer for error processing where it would be recognized as a BIER Echo Request packet. BFR B MUST send BIER Echo Reply to BFIR A and MUST include Downstream Mapping TLV, defined in [I-D.ietf-bier-ping] setting its fields in the following fashion:

- MTU SHOULD be set to the minimal MTU value among all BIER-enabled egress interfaces toward downstream BFRs that could be used to reach B's downstream BFERs;

- Address Type MAY be set to any value defined in Section 3.3.4 [I-D.ietf-bier-ping].

- I flag MUST be cleared to direct the responding BFR not to include the Incoming SI-BitString TLV in the BIER Echo Response.

- Downstream Interface Address field MUST be zeroed.

- List of Sub-TLVs MUST include the Egress Bitstring sub-TLV with the list of all BFERs that cannot be reached because the egress MTU turned out to be too small.

The BFIR will receive either of the two types of packets:

- a positive Echo Reply from one of BFERs to which the probe has been sent. In this case, the bit corresponding to the BFER MUST be cleared from the bitmask string (BMS);

- a negative Echo Reply with bit string listing unreached BFERs and recommended MTU value MTU'. The BFIR MUST add the bit string to its BMS and set the size of the next probe as min(MTU, MTU')

If a negative Echo Reply is received, the BFIR MUST wait for the expiration of the Echo Request before transmitting the updated Echo Request. If upon expiration of the Echo Request timer BFIR didn't receive any Echo Replies, then the size of the probe SHOULD be decreased. There are scenarios when an implementation of the PMTUD would not decrease the size of the probe. For example, suppose upon expiration of the Echo Request timer BFIR didn't receive any Echo Reply. In that case, BFIR MAY continue to retransmit the probe using the initial size and MAY apply probe delay retransmission procedures. The algorithm used to delay retransmission procedures on BFIR is outside the scope of this specification. The BFIR sends probes using BMS and locally defined retransmission procedures, but not more frequently than after the Echo Request timer expired, until either the bit string is clear, i.e., contains no set bits, or until the BFIR retransmission procedure terminates and PMTU discovery is declared unsuccessful. In the case of convergence of the procedure, the size of the last probe indicates the PMTU size that can be used for all BFERs in the initial BMS without incurring fragmentation.

Thus we conclude that in order to comply with the requirement in [I-D.ietf-bier-oam-requirements]:

- a BFR SHOULD support PMTUD;

- a BFR MAY use defined per BIER sub-domain MTU value as initial MTU value for discovery or use it as MTU for this BIER sub-domain to reach BFERs;

- a BFIR MUST have a locally defined PMTUD probe retransmission procedure.

3.1. Data TLV for BIER Ping

There needs to be a control for probe size in order to support the BIER PMTUD. Data TLV format is presented in Figure 2. Data TLV MAY be added in BIER Echo Request or Echo Reply message as defined in [I-D.ietf-bier-ping].

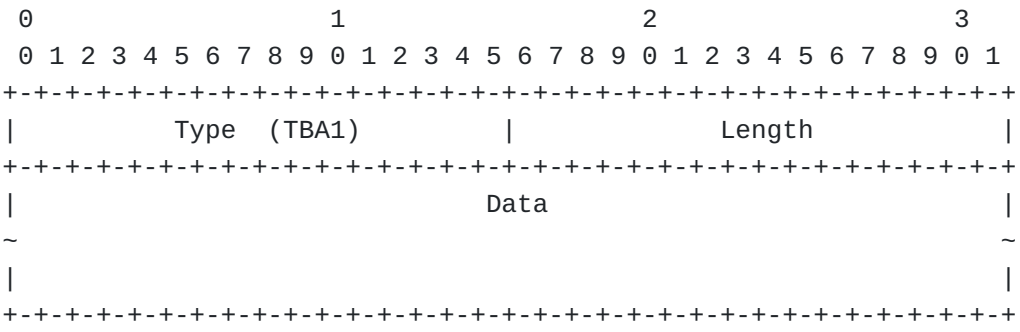

Figure 2: Data TLV format

- Type: indicates Data TLV, to be allocated by IANA Section 4.

- Length: the length of the Data field in octets.

- Data: n octets (n = Length) of arbitrary data. The receiver SHOULD ignore it.

4. IANA Considerations

IANA is requested to assign a new Type value for Data TLV Type from its registry of TLV and sub-TLV Types of BIER Ping as follows:

Value	Description	Reference
TBA1	Data	This document

Table 1: Data TLV Type

5. Security Considerations

Routers that support PMTUD based on this document are subject to the same security considerations as defined in [I-D.ietf-bier-ping]

6. Acknowledgment

Authors greatly appreciate thorough review and the most detailed comments by Eric Gray.

7. References

7.1. Normative References

[I-D.ietf-bier-ping] Nainar, N. K., Pignataro, C., Chen, M., and G. Mirsky, "BIER Ping and Trace", Work in Progress, Internet-Draft, draft-ietf-bier-ping-12, 29 July 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bier-ping-12>.

[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>.

[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>.

[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>.

[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>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

7.2. Informative References

[I-D.ietf-bier-oam-requirements]

Mirsky, G., Nainar, N. K., Chen, M., and S. Pallagatti, "Operations, Administration and Maintenance (OAM) Requirements for Bit Index Explicit Replication (BIER) Layer", Work in Progress, Internet-Draft, draft-ietf-bier-oam-requirements-13, 10 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bier-oam-requirements-13>.

Authors' Addresses

Greg Mirsky
Ericsson

Email: gregimirsky@gmail.com

Tony Przygienda
Juniper Networks

Email: prz@juniper.net

Andrew Dolganow
Nokia

Email: andrew.dolganow@nokia.com