BIER Working Group   G. Mirsky
Internet-Draft   Ericsson
Intended status: Standards Track   T. Przygienda
Expires: October 7, 2016   Juniper Networks
                                                         A. Dolganow
                                                               Nokia
                                                       April 5, 2016

Path Maximum Transmission Unit Discovery (PMTUD) for Bit Index Explicit
                      Replication (BIER) Layer
                draft-mirsky-bier-path-mtu-discovery-01

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 http://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 October 7, 2016.

Copyright Notice

Table of Contents

1. Introduction

In packet switched networks when a host seeks to transmit a sizable
amount of data to a target destination the data is transmitted as a
set of datagrams. In most cases it is more efficient to use the
largest possible datagrams but so that these datagrams do not have to
be fragmented at any point along the path from the host to the
destination in order to avoid performance degradation caused by
fragmentation. Fragmentation occurs on hops along the route where an
Maximum Transmission Unit (MTU) is smaller than the size of the
datagram. To avoid such fragmentation the MTU for each hop along a
path from a host to a destination must be known to select an
appropriate datagram size. Such MTU determination along a specific
path is referred to as path MTU discovery (PMTUD).

[I-D.ietf-bier-architecture] introduces and explains Bit Index
Explicit Replication (BIER) architecture and how it supports
forwarding of multicast data packets. A BIER domain consists of Bit-
Forwarding Routers (BFRs) that are uniquely identified by their
respective BFR-ids. An ingress border router (acting as a Bit
Forwarding Ingress Router (BFIR)) inserts a Forwarding Bit Mask
(F-BM) into a packet. Each targeted egress node (referred to as a
Bit Forwarding Egress Router (BFER)) is represented by Bit Mask
Position (BMP) in the BMS. A transit or intermediate BIER node,
referred as BFR, forwards BIER encapsulated packets to BFERs,
identified by respective BMPs, according to a Bit Index Forwarding
Table (BIFT).

1.1. Conventions used in this document

1.1.1. Terminology

BFR: Bit-Forwarding Router

BFER: Bit-Forwarding Egress Router

BFIR: Bit-Forwarding Ingress Router

BIER: Bit Index Explicit Replication

BIFT: Bit Index Forwarding Tree

F-BM: Forwarding Bit Mask

MTU: Maximum Transmission Unit

OAM: Operations, Administration and Maintenance

PMTUD: Path MTU Discovery

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
[RFC2119].

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.kumarzheng-bier-ping] for use in BIER PMTUD
solution.

Current PMTUD mechanisms [RFC1191], [RFC1981], 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 result, a transient node that
cannot forward a probe packet that is bigger than its link MTU sends
to the ingress node an error notification, otherwise the egress
responds with a positive acknowledgement. Thus, through series of
iterations, decreasing and increasing size of the probe packet, the
ingress node discovers the MTU of the particular path.

Thus applied 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
over and over again until the very last egress responds with a
positive acknowledgement. Consider without loss of generality an
example multicast network presented in Figure 1, where MTU on all
links but one (B,D) is the same. If MTU on link (B,D) is smaller
than the MTU on the other links, using existing PMTUD mechanism
probes will unnecessary flood to leaf nodes E, F, and G for the
second and consecutive times and positive responses will be generated
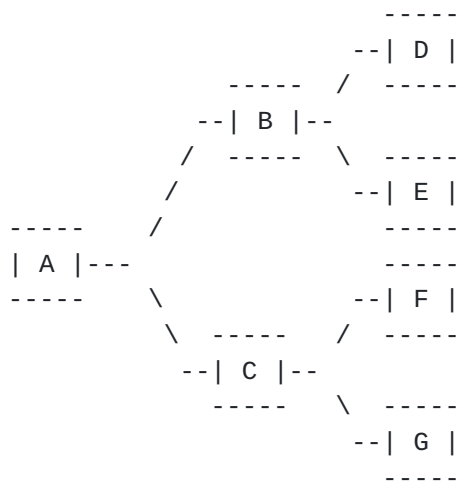and received by root A repeatedly.

```
                            -----
                         --| D |
                -----   /  -----
              --| B |--
             /  -----  \  -----
            /           --| E |
   -----   /              -----
  | A |---                -----
   -----   \              --| F |
            \  -----    /  -----
             --| C |--
                -----   \  -----
                         --| G |
                            -----
```

Figure 1: Multicast network

3. PMTUD Mechanism for BIER

A BFIR selects a set of BFERs for the specific multicast
distribution. Such BFIR determines, by explicitly controlling subset
of targeted BFERs and transmitting series of probe packets, the MTU
of that multicast distribution path. The critical step is that in
case of failure at an intermediate BFR to forward towards the subset
of targeted downstream BFERs, the BFR responds with a partial
(compared to the one it received in the request) bitmask towards the
originating BFIR in error notification. That allows for
retransmission of the next probe with smaller MTU address only
towards the failed downstream BFERs instead of all BFERs addressed in
the previous probe. In the scenario discussed in Section 2 the
second and all following (if needed) probes will be sent only to the
node D since MTU discovery of E, F, and G has been completed already
by the first probe successfully.

[I-D.kumarzheng-bier-ping] introduced BIER Ping as transport-
independent OAM mechanism to detect and localize failures in BIER

data plane. This document specifies how BIER Ping can be used to
perform efficient PMTUD in BIER domain.

Consider network displayed in Figure 1 to be 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 been assumed in
Section 2, MTUs of all links but 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 BIER Echo
Request packet. BFR B MUST send BIER Echo Reply to BFIR A and MUST
include Downstream Mapping TLV, defined in [I-D.kumarzheng-bier-ping]
setting its fields in the following fashion:

o MTU SHOULD be set to minimal MTU value among all egress BIER links
   that could be used to reach B's downstream BFERs;

o Address Type MUST be set to 0 [Ed.note: we need to define 0 as
   valid value for the Address Type field with the specific semantics
   to "Ignore" it.]

o I flag MUST be cleared;

o Downstream Interface Address field (4 octets) MUST be zeroed and
   MUST include in Egress Bitstring sub-TLV the list of all BFERs
   that cannot be reached because the attempted MTU turned out to be
   too small.

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

o a positive Echo Reply from one of BFERs to which the probe has
   been sent. In such case the bit corresponding to the BFER MUST be
   cleared from the BMS;

o 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 size of the next probe as min(MTU, MTU")

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, if upon expiration of the Echo
Request timer BFIR didn't receive any Echo Reply, then 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 MUST continue sending probes using BMS until
the bit string is clear or the discovery is declared unsuccessful.
In case of convergence of the procedure, the size of the last probe
indicates the MTU 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]:

o a BFR SHOULD support PMTUD;

o 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.

3.1. Data TLV for BIER Ping

There need to be control of probe size in order to support the BIER
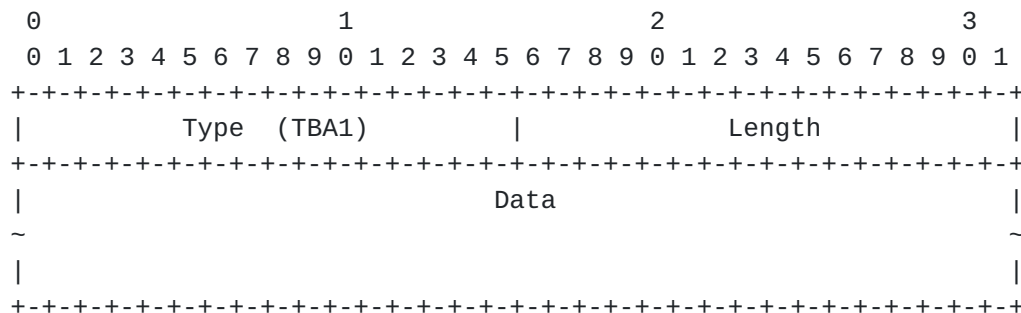PMTUD. Data TLV format is presented in Figure 2.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type  (TBA1)         |              Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Data                              |
~                                                               ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Data TLV format

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

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

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

4. IANA Considerations

IANA is requested to assign 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.kumarzheng-bier-ping]

6. Acknowledgement

   TBD

7. References

7.1. Normative References

   [I-D.ietf-bier-architecture]
             Wijnands, I., Rosen, E., Dolganow, A., P, T., and S.
             Aldrin, "Multicast using Bit Index Explicit Replication",
             draft-ietf-bier-architecture-03 (work in progress),
             January 2016.

   [I-D.kumarzheng-bier-ping]
             Kumar, N., Pignataro, C., Akiya, N., Zheng, L., Chen, M.,
             and G. Mirsky, "BIER Ping and Trace", draft-kumarzheng-
             bier-ping-02 (work in progress), December 2015.

   [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191,
             DOI 10.17487/RFC1191, November 1990,
             <http://www.rfc-editor.org/info/rfc1191>.

   [RFC1981] McCann, J., Deering, S., and J. Mogul, "Path MTU Discovery
             for IP version 6", RFC 1981, DOI 10.17487/RFC1981, August
             1996, <http://www.rfc-editor.org/info/rfc1981>.

   [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
             Requirement Levels", BCP 14, RFC 2119,
             DOI 10.17487/RFC2119, March 1997,
             <http://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,
             <http://www.rfc-editor.org/info/rfc4821>.

[7.2](#). **Informative References**

   [I-D.ietf-bier-oam-requirements]
              Mirsky, G., Nordmark, E., Pignataro, C., Kumar, N.,
              Aldrin, S., Zheng, L., Chen, M., Akiya, N., and S.
              Pallagatti, "Operations, Administration and Maintenance
              (OAM) Requirements for Bit Index Explicit Replication
              (BIER) Layer", [draft-ietf-bier-oam-requirements-01](#) (work
              in progress), March 2016.

Authors' Addresses

   Greg Mirsky
   Ericsson

   Email: gregory.mirsky@ericsson.com

   Tony Przygienda
   Juniper Networks

   Email: prz@juniper.net

   Andrew Dolganow
   Nokia

   Email: andrew.dolganow@nokia.com