MTU Signalling Extensions for LDP
draft-ietf-mpls-ldp-mtu-extensions-00

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on December 30, 2002.

Copyright Notice

Copyright (C) The Internet Society (2002). All Rights Reserved.

Abstract

Proper functioning of RFC 1191 path MTU detection requires that IP routers have knowledge of the MTU for each link to which they are connected. As currently specified, LDP does not have the ability to signal the MTU for an LSP to ingress LSRs. In the absence of this functionality, the MTU for each LSP must be statically configured by network operators or by equivalent, off-line mechanisms.

This document specifies extensions to the LDP label distribution protocol in support of LSP MTU signalling.
1. Introduction

As currently specified in [3], the LDP protocol for MPLS does not support signalling of the MTU for LSPs to ingress LSRs. This functionality is essential to the proper functioning of RFC 1191 path MTU detection [1]. Without knowledge of the MTU for an LSP, edge LSRs may transmit packets along that LSP which are, according to [4], too big. Such packets may be silently discarded by LSRs along the LSP, effectively preventing communication between certain end hosts.

The solution proposed in this document enables automatic determination of the MTU for an LSP with the addition of a TLV to carry MTU information for a FEC between adjacent LSRs in LDP Label Mapping messages. This information is sufficient for a set of LSRs along the path followed by an LSP to discover either the exact MTU for that LSP, or an approximation which is no worse than could be generated with local information on the ingress LSR.
2. MTU Signalling

The signalling procedure described in this document employs the addition of a single TLV to LDP Label Mapping messages and a simple algorithm for LSP MTU calculation.

2.1 Signalling Procedure

The procedure for signalling the MTU is performed hop-by-hop by each LSR L along an LSP. The steps are as follows:

1. First, L computes the MTU for each FEC:
   1. If L is the egress LSR for the FEC, L set the MTU to the MTU of the egress interface, unless local policy specifies otherwise.
   2. If L is not the egress LSR for the FEC, L SHOULD set the MTU to 0xffff, indicating that it is not the egress LSR and has not yet received an MTU other than 0xffff from downstream LSRs. Local policy may dictate the selection of a value other than 0xffff, but the default in the absence of such policy should be 0xffff.
   3. If L is not the egress LSR for a FEC, and L receives a Mapping for a FEC which includes an MTU TLV with a value other than 0xffff, L calculates the MTU according to the rules in Section 2.2. If L receives multiple Mapping messages for this FEC, it first chooses between them by some policy, then applies the calculation for the chosen Mapping. This is the "active Mapping" for this FEC.
   4. If L receives a Mapping for a FEC without an MTU TLV from a directly connected neighbor, L MAY act as if it received an MTU TLV with MTU 0xffff, and follow the procedure in Step 1.2. Otherwise, L MUST send Mappings for this FEC without an MTU TLV.
   5. If L receives a Mapping for a FEC from a peer to which it is not directly connected, it must first find an LSP by which L can reach the peer. (Note that this procedure may be recursively applied.) Once the appropriate LSP has been determined, the MTU is calculated according to the rules in Section 2.2, using the MTU of the selected LSP as the link MTU.

2. For each direct LDP neighbor of L to which L decides to send a Mapping for a FEC, L attaches an MTU TLV with the MTU that it
computed for this FEC. Mapping messages sent to "remote" LDP neighbors need not have an MTU TLV.

3. When a new MTU is received for a label mapping from a downstream LSR, or the active Mapping for a FEC changes, L returns to Step 1. If the newly computed MTU is unchanged, L does not advertise new information to its neighbors.

This behavior is standard for attributes such as path vector and hop count, and the same rules apply, as specified in [3].

4. In some cases, a node may act as both an LER and an LSR for the same LSP. In these situations, the node will calculate multiple MTUs: the MTU advertised to upstream LSRs for labelled traffic and the MTU used locally when processing unlabelled traffic. The procedure for calculating each of these MTUs is unchanged from the steps above, although the series of steps taken will differ depending on which MTU is being calculated.

2.2 Calculating Local MTU

There is a wide variety of policies which may be used in determining the MTU advertised by a node, however there are restrictions which MUST be adhered to in order to ensure proper operation of MTU signalling and minimization of signalling traffic during topology changes.

If the local policy is based entirely on the egress interface for the LSP, the calculated MTU must be less than or equal to the egress interface MTU.

If the local policy is based on a group of egress interfaces, the calculated MTU MUST be less than or equal to the MTU of the egress interface with the largest MTU in the group minus any label overhead, but SHOULD be less than or equal to the MTU of the egress interface with the smallest MTU in the group minus any label overhead.

If the local LSR is the ingress LER for the FEC in question, any label overhead introduced must be subtracted from the calculated MTU to determine the actual path MTU. For example, if 2 labels are pushed onto the stack, the LSR MUST subtract 8 bytes from the MTU value it has calculated based on local link MTUs and MTU values received from downstream LDP neighbors.

Under no circumstances must the advertised MTU exceed the received MTU.
2.3 MTU TLV

The MTU TLV encodes information on the maximum transmission unit for an LSP, either for the entire path or only for a segment of the path.

The encoding for the MTU TLV is:

```
0                   1                   2                   3
+-----------------------------+-----------------------------+
|1|0| MTU TLV (0x0XXX) | Length |
+-----------------------------+-----------------------------+
| MTU |                          |
+-----------------------------+
```

MTU

This is a 16-bit unsigned integer that represents the MTU in bytes for an LSP or segment of an LSP.
3. Example of Operation

The figure and below describes a simple LSR topology. Ri and Re are the ingress and egress LSRs for LSP P1. Rx and Re are the ingress and egress LSRs for LSP P2. From Rx to Re, LSP P1 is encapsulated in LSP P2. Ry is an intermediate LSR which does not act as ingress or egress for any LSPs. L1 through L3 are links connecting the LSRs. Le is the egress link.

```
+---+       +---+       +---+       +---+       Link   MTU    overhead
| Ri|--L1--| Rx|--L2--| Ry|--L3--| Re|--Le   ----  ------  --------
+---+       +---+       +---+       +---+        L1    9216     9216
 |         |                  ^^         L2    4470     4466
 |         |                  ||         L3    9216     9212
 |         +---P2-------------+|         Le    9216     9216
 +-------------P1--------------+
```

Figure 1. Sample LSR Topology

The following four time steps illustrate the calculation of the MTU for P1. Let FEC F represent traffic mapped to LSP P1.

At t[0]:

1) Re sets the MTU for F to 9216 (the MTU of the egress interface) and sends a Mapping message for F to Ry.

2) Ri, Rx, and Ry have not received Mappings for F.

At t[1]:

1) Ry receives a Mapping for F from Re with an MTU of 9216. Ry compares 9216 to 9216 (Ry does not push a label onto the stack for either P1 or P2), and sends a mapping message for F with an MTU of 9216 to Rx.

2) Ri and Rx have not received Mappings for F.

At t[2]:

1) Rx receives a Mapping for F from Ry with an MTU of 9216. Rx compares 9212 (9216 - 4) to 4466, and sends a Mapping message for F with an MTU of 4466 to Ri.
2) Ri has not received Mappings for F.

At t[3]:

1) Ri receives a Mapping for F from Rx with an MTU of 4462. Ri compares 4466 to 9216, and sets the MTU for P1 to 4462 (4466 minus the overhead of 1 label pushed onto the stack).
4. Protocol Interaction

4.1 Interaction With LSRs Which Do Not Support MTU Signalling

Changes in MTU for sections of an LSP may cause intermediate LSRs to generate unsolicited label Mapping messages to advertise the new MTU. LSRs which do not support MTU signalling MUST accept these messages, but MAY ignore them (see Section 2.1).

4.2 Interaction with CR-LDP and RSVP-TE

The MTU TLV can be used to discover the Path MTU of both LDP LSPs and CR-LDP LSPs. This proposal is not impacted in the presence of LSPs created using CR-LDP, as specified in [2].

Note that LDP/CR-LDP LSPs may tunnel through other LSPs signalled using LDP, CR-LDP or RSVP-TE [5]; the mechanism suggested here applies in all these cases.
5. Security Considerations

This mechanism does not introduce any new weaknesses in LDP. It is possible to spoof TCP packets belonging to an LDP session to manipulate the LSP MTU, but this sort of attack is not new to LDP.
6. Acknowledgments

We would like to thank Andre Fredette for a number of detailed comments on earlier versions of the signalling mechanism. Eric Gray and Giles Heron have contributed numerous useful suggestions.
References (Normative)


Authors’ Addresses

Benjamin Black
Layer8 Networks
EMail: ben@layer8.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA  94089
US
EMail: kireeti@juniper.net