Network Working Group
Request for Comments: 3988
Category: Experimental

B. Black
Layer8 Networks
K. Kompella
Juniper Networks

January 2005

Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol

Status of This Memo

Copyright Notice

Abstract

Proper functioning of RFC 1191 path Maximum Transmission Unit (MTU) discovery requires that IP routers have knowledge of the MTU for each link to which they are connected. As currently specified, the Label Distribution Protocol (LDP) does not have the ability to signal the MTU for a Label Switched Path (LSP) to the ingress Label Switching Router (LSR). 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 experimental extensions to LDP in support of LSP MTU discovery.

1. Introduction

As currently specified in [2], 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 [3]. Without knowledge of the MTU for an LSP, edge LSRs may transmit packets along that LSP which are, according to [4], too big. These 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 by adding a Type-Length-Value triplet (TLV) to carry MTU information for a Forwarding Equivalence Class (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 that is no worse than could be generated with local information on the ingress LSR.

1.1. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14, RFC 2119 [1].

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. Definitions

Link MTU: The MTU of a given link. This size includes the IP header and data (or other payload) and the label stack but does not include any lower-layer headers. A link may be an interface (such as Ethernet or Packet-over-SONET), a tunnel (such as GRE or IPsec), or an LSP.

Peer LSRs: For LSR A and FEC F, this is the set of LSRs that sent a Label Mapping for FEC F to A.

Downstream LSRs: For LSR A and FEC F, this is the subset of A’s peer LSRs for FEC F to which A will forward packets for the FEC. Typically, this subset is determined via the routing table.

Hop MTU: The MTU of an LSP hop between an upstream LSR, A, and a downstream LSR, B. This size includes the IP header and data (or other payload) and the part of the label stack that is considered payload as far as this LSP goes. It does not include any lower-level headers. (Note: If there are multiple links between A and B, the Hop MTU is the minimum of the Hop MTU of those links used for forwarding.)

LSP MTU: The MTU of an LSP from a given LSR to the egress(es), over each valid (forwarding) path. This size includes the IP header and data (or other payload) and any part of the label stack that was received by the ingress LSR before it placed the packet into the LSP (this part of the label stack is considered part of the payload for this LSP). The size does not include any lower-level headers.

2.2. Example

Consider LSRs A - F, interconnected as follows:

```
                     M       P
                   _____ C =====
                  /      |      \
         A ~~~~~ B ===== D ----- E ----- F
             L       N       Q       R
```

Say that the link MTU for link L is 9216; for links M, Q, and R, 4470; and for N and P, is 1500.

Consider an FEC X for which F is the egress, and say that all LSRs advertise X to their neighbors.

Note that although LDP may be running on the C - D link, it is not used for forwarding (e.g., because it has a high metric). In particular, D is an LDP neighbor of C, but D is not one of C’s downstream LSRs for FEC X.

E’s peers for FEC X are C, D, and F. Say that E chooses F as its downstream LSR for X. E’s Hop MTU for link R is 4466. If F advertised an implicit null label to E, then E MAY set the Hop MTU for R to 4470.

C’s peers for FEC X are B, D, and E. Say that C chooses E as its downstream LSR for X. Similarly, A chooses B, B chooses C and D (equal cost multi-path), D chooses E, and E chooses F (respectively) as downstream LSRs.

C’s Hop MTU to E for FEC X is 1496. B’s Hop MTU to C is 4466 and to D is 1496. A’s LSP MTU for FEC X is 1496. If A has another LSP for FEC Y to F (learned via targeted LDP) that rides over the LSP for FEC X, the MTU for that LSP would be 1492.

If B had a targeted LDP session to E (e.g., over an RSVP-TE tunnel T) and B received a Mapping for FEC X over the targeted LDP session, then E would also be B’s peer, and E may be chosen as a downstream LSR for B. In that case, B’s LSP MTU for FEC X would then be the smaller of {(T’s MTU - 4), E’s LSP MTU for X}.

This memo describes how A determines its LSP MTU for FECs X and Y.

2.3. Signalling Procedure

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

1. First, L computes its LSP MTU for FEC F:

   A. If L is the egress for F, L sets the LSP MTU for F to 65535.

   B. [OPTIONAL] If L’s only downstream LSR is the egress for F (i.e., L is a penultimate hop for F) and L receives an implicit null label as its Mapping for F, then L can set the Hop MTU for its downstream link to the link MTU instead of (link MTU - 4 octets). L’s LSP MTU for F is the Hop MTU.

   C. Otherwise (L is not the egress LSR), L computes the LSP MTU for F as follows:

      a) L determines its downstream LSRs for FEC F.

      b) For each downstream LSR Z, L computes the minimum of the Hop MTU to Z and the LSP MTU in the MTU TLV that Z advertised to L. If Z did not include the MTU TLV in its Label Mapping, then Z’s LSP MTU is set to 65535.

      c) L sets its LSP MTU to the minimum of the MTUs it computed for its downstream LSRs.

2. For each LDP neighbor (direct or targeted) of L to which L decides to send a Mapping for FEC F, L attaches an MTU TLV with the LSP MTU that it computed for this FEC. L MAY (because of policy or for other reasons) advertise a smaller MTU than it has computed, but L MUST NOT advertise a larger MTU.

3. When a new MTU is received for FEC F from a downstream LSR or the set of downstream LSRs for F changes, L returns to step 1. If the newly computed LSP MTU is unchanged, L SHOULD NOT advertise new information to its neighbors. Otherwise, L readvertises its Mappings for F to all its peers with an updated MTU TLV.

   This behavior is standard for attributes such as path vector and hop count, and the same rules apply, as specified in [2].

   If the LSP MTU decreases, L SHOULD readvertise the new MTU immediately; if the LSP MTU increases, L MAY hold down the readvertisement.

2.4. MTU TLV

The MTU TLV encodes information on the maximum transmission unit for an LSP, from the advertising LSR to the egress(es) over all valid paths.

The encoding for the MTU TLV is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|1|      MTU TLV (0x0601)     |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              MTU              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

MTU

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

Note that the U and F bits are set. An LSR that doesn’t recognize the MTU TLV MUST ignore it when it processes the Label Mapping message and forward the TLV to its peers. This may result in the incorrect computation of the LSP MTU; however, silently forwarding the MTU TLV preserves the maximal amount of information about the LSP MTU.

3. Example of Operation

Consider the network example in Section 2.2. For each LSR, Table 1 describes the links to its downstream LSRs, the Hop MTU for the peer, the LSP MTU received from the peer, and the LSR’s computed LSP MTU.

Now consider the same network with the following changes: There is an LSP T from B to E, and a targeted LDP session from B to E. B’s peer LSRs are A, C, D, and E; B’s downstream LSRs are D and E; to reach E, B chooses to go over T. The LSP MTU for LSP T is 1496. This information is depicted in Table 2.

LSR	Link	Hop MTU	Recvd MTU	LSP MTU
F	-	65535	-	65535
E	R	4466	F: 65535	4466
D	Q	4466	E: 4466	4466
C	P	1496	E: 4466	1496
B	M	4466	C: 1496	
	N	1496	D: 4466	1496
A	L	9212	B: 1496	1496

Table 1

LSR	Link	Hop MTU	Recvd MTU	LSP MTU
F	-	65535	-	65535
E	R	4466	F: 65535	4466
D	Q	4466	E: 4466	4466
C	P	1496	E: 4466	1496
B	T	1492	E: 4466	
	N	1496	D: 4466	1492
A	L	9212	B: 1492	1492

Table 2

4. Using the LSP MTU

An ingress LSR that forwards an IP packet into an LSP whose MTU it knows MUST either fragment the IP packet to the LSP's MTU (if the Don't Fragment bit is clear) or drop the packet and respond with an ICMP Destination Unreachable message to the source of the packet, with the Code indicating "fragmentation needed and DF set", and the Next-Hop MTU set to the LSP MTU. In other words, the LSR behaves as RFC 1191 says, except that it treats the LSP as the next hop "network".

If the payload for the LSP is not an IP packet, the LSR MUST forward the packet if it fits (size <= LSP MTU) and SHOULD drop it if it doesn't.

5. Protocol Interaction

5.1. Interaction with LSRs that 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 that do not support MTU signalling will, due to message and TLV processing mechanisms specified in RFC3036 [2], accept the messages carrying the MTU TLV but will ignore the TLV and forward the TLV to the upstream nodes (see Section 2.4).

5.2. 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 with CR-LDP, as specified in [5].

Note that LDP/CR-LDP LSPs may tunnel through other LSPs signalled using LDP, CR-LDP, or RSVP-TE [6]; the mechanism suggested here applies in all of these cases, essentially by treating the tunnel LSPs as links.

6. References

6.1. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[2] Andersson, L., Doolan, P., Feldman, N., Fredette, A., and B. Thomas, "LDP Specification", RFC 3036, January 2001.

[3] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, November 1990.

[4] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

[6] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

6.2. Informative References

[5] Jamoussi, B., Andersson, L., Callon, R., Dantu, R., Wu, L., Doolan, P., Worster, T., Feldman, N., Fredette, A., Girish, M., Gray, E., Heinanen, J., Kilty, T., and A. Malis, "Constraint-Based LSP Setup using LDP", RFC 3212, January 2002.

7. 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 LDP has mechanisms to thwart these types of attacks. See Section 5 of [2] for more information on security aspects of LDP.

8. IANA Considerations

IANA has allocated 0x0601 as a new LDP TLV Type, defined in Section 2.4. See: http://www.iana.org/assignments/ldp-namespaces

9. Acknowledgements

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

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

Full Copyright Statement

Intellectual Property

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.