Network Working Group
Internet Draft
Document: draft-galtzur-l2tpext-gr-02.txt

S. Galtzur
Axerra Networks

Expires: April 2005
October 2004

Layer Two Tunneling Protocol (Version 3) Graceful Restart

draft-galtzur-l2tpext-gr-02.txt

Status of this Memo

Abstract

This document describes a mechanism that helps to minimize the negative effects on L2TP traffic caused by L2TP Control Connection Endpoint (LCCE) control plane restart, specifically by the restart of its control protocol component, on LCCEs that are capable of preserving the L2TP forwarding component (a.k.a. the L2TP data plane) across the restart.

The mechanism described in this document is applicable to all LCCEs, both those with the ability to preserve Forwarding State during the control plane (CP) restart and those without (although the latter needs to implement only a subset of the mechanism described in this document).
Supporting (a subset of) the mechanism described here by the LCCEs that can not preserve their L2TP Forwarding State across the restart would not reduce the negative impact on L2TP traffic caused by their control plane restart, but it would minimize the impact on the L2TP traffic if their peer(s) are capable of preserving the Forwarding State across the restart of their control plane and implement the mechanism described here.

The mechanism makes minimalistic assumptions on what has to be preserved across restart - the mechanism assumes that only the actual L2TP Forwarding State has to be preserved; the mechanism does not require any of the control plane related states to be preserved across the restart.

Conventions used in this document

For the sake of brevity in the context of this document, by "the control plane" we mean "the L2TP component of the control plane". The L2TP control plane includes all the information associated with the L2TP Control Connection and the low-order reliable delivery protocol.

For the sake of brevity in the context of this document, by "L2TP Forwarding State" we mean the dynamic information that is exchanged between two LCCEs peers during the establishment of L2TP tunnels and sessions, i.e. local and remote Session IDs and local and remote cookies. The Forwarding State of an L2TP session also includes its association with the specific end service or application.
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 RFC2119 [2].

Table of Contents

1. Motivation

The mechanism presented in this document extends the ideas first explored in [4] for LDP graceful restart to L2TPv3. L2TPv3 [3] is the protocol of choice for setup, teardown and maintenance of pseudo-wires over an IP PSN (see [6]) just as LDP is the protocol of choice for setup, teardown and maintenance of pseudo-wires over an MPLS PSN [7], with the PWE3 Provider Edge (PE) devices acting also as L2TPv3 Control Connection Endpoints (LCCEs).

In the case where a LCCE could preserve its L2TP Forwarding State across restart of its control plane, it is desirable not to perturb the L2TP Session IDs going through that LCCE. In this document, we describe a mechanism, termed "L2TP Graceful Restart", that allows the accomplishment of this goal.

The mechanism described in this document is applicable to all LCCEs, both those with the ability to preserve Forwarding State during control plane restart and those without (although the latter need to implement only a subset of the mechanism described in this document). Supporting (a subset of) the mechanism described here by the LCCEs that can not preserve their L2TP Forwarding State across the restart would not reduce the negative impact on L2TP traffic caused by their control plane restart, but it would minimize the impact if their peer(s) are capable of preserving the Forwarding State across the restart of their control plane and implement the mechanism described here.

The mechanism makes minimalistic assumptions on what has to be preserved across restart - the mechanism assumes that only the actual L2TP Forwarding State has to be preserved. Clearly this is the minimum amount of state that has to be preserved across the restart in order not to perturb the L2TP Session IDs terminating in a restarting LCCE. The mechanism does not require any of the L2TP-related states to be preserved across the restart.

2. Changes from the Previous Version

1. Processing of non-established sessions by the peer of the restarting LCCE has been clarified.

2. The list of control messages that can use the Session Graceful Restart AVP has been updated.

3. Lack of additional security risks of the Graceful Restart mechanism has been explained.

3. L2TP Extension

There is one new AVP for the Control Connection Messages and one new AVP for the Session Connection Messages. There is also one new state in the Session state machine and one new Error Code value.

3.1 Graceful Restart AVP [SCCRQ, SCCRP]

An LCCE that supports (fully or partially) L2TP Graceful Restart as defined in this document MUST include the Graceful Restart (GR) AVP in the SCCRQ and SCCRP messages.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |M|H| rsvd  |      Length       |           Vendor Id [IETF]    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Attribute Type [TBD]  |           Reserved            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              GR Reconnect Timeout (in milliseconds)           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              Recovery Time (in milliseconds)                  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

This AVP MAY be hidden (the H bit MAY be 0 or 1). The M bit for this AVP SHOULD be set to 0. The Length (before hiding) of this AVP is 16.

The GR Reconnect Timeout is the time (in milliseconds) the initiating LCCE asks its peer to wait after the next detection of communication failure for a Graceful Restart Reconnection. Value of zero indicates that the LCCE does not preserve its L2TP Forwarding State across the restart of the L2TP control plane, so that the peer should not wait for a graceful restart of this LCCE.

The Recovery Time is the time (in milliseconds) the initiating LCCE asks its peer to wait after the establishment of this control connection for recovery of the Sessions that belong to this Control

Connection. Value of zero indicates that the sending LCCE was not able to preserve the Forwarding State and restart as described in [3] should be used.

3.2 Graceful Restart Session AVP [ICRQ, OCRQ, ICRP, OCRP]

An LCCE that tries to open a session for which the L2TP Forwarding State has been preserved, MUST include the Graceful Restart Session AVP when trying to reopen the Session gracefully.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |M|H| rsvd  |      Length       |           Vendor Id [IETF]    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Attribute Type [TBD]  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

When this AVP is present in an ICRQ/OCRQ/ICRP/OCRP message the value of the Remote Session ID in the Remote Session ID AVP MUST be set to the preserved value of the Remote Session ID.

This AVP MAY be hidden (the H bit MAY be 0 or 1). The M bit for this AVP SHOULD be set to 0. The Length (before hiding) of this AVP is 6.

3.3 Stale state in the Session state machine

A Session enters the stale state if it has been in the established state and its associated Control Connection enters the Graceful Restart procedure as described in the following section. Forwarding of L2TP data packets for a Session in this state remains unperturbed.

A Session transits from the stale state to either the established state or the idle state as described in the following section.

3.4 A New Error Code value

One new Error Code value - Session Graceful Restart Mismatch (actual value to be assigned by IANA) will be used in the CDN messages with the Result Code 2 (Session disconnected for the reason indicated in Error Code) as defined in [3] in the following situations:

- o Attempt to re-establish a non-stale session with a Session initiation request that contains the Session Graceful Restart AVP

- o Attempt to re-establish a stale session with a Session initiation request that does not contains the Session Graceful Restart AVP.

4. Operation

An LCCE that support the Graceful Restart, as defined in this document, advertises it by including the GR AVP in the SCCRQ or SCCRP Messages. If one of the peers does not include this AVP both LCCEs MUST follow the Control Connection initiation procedure as described in [3].

4.1 Procedure for restarting LCCE

After an LCCE restarts its control plane, it MUST check whether it was able to preserve its L2TP Forwarding State across the restart. If not, then the LCCE sets the Recovery Time to 0 in the GR AVP it sends to its peer when the Control Connection is re-established. If the L2TP Forwarding State has been preserved, the LCCE starts its internal timer, called L2TP Forwarding State Holding timer (the value of that timer SHOULD be configurable and MUST NOT be greater then the GR Reconnect Timeout sent on the previous GR AVP), and all the established L2TP Sessions transit to the stale state.

Note: all the sessions that are not in the established state MUST transit to the idle state since they will never be recovered by the Graceful Restart mechanism.

While this procedure is performed the LCCE MUST ignore any incoming SCCRQ messages for the Control Connections being recovered. At the expiration of the timer, all the entries still in the stale state MUST transit to the idle state (see [3] for state machine details). The value of the Recovery Time advertised in the GR AVP is set to the (current) value of the timer at the point in which the Initiation message carrying the GR AVP is sent.

We say that an LCCE is in the process of restarting when the L2TP Forwarding State Holding timer has not expired. Once the timer expires, we say that the LCCE has completed its restart.

When the LCCE receives the GR AVP from its peer it MUST set its L2TP Forwarding State Holding timer to the smaller value of the its own current value and the Recovery Time as advertised by the peer.

4.2 Restart of L2TP communication with a peer LCCE

When an LCCE detect that its L2TP Control Connection with its peer LCCE went down, and the LCCE knows that the peer is capable of preserving its L2TP Forwarding State across the restart (as was indicated by the presence of GR AVP with non-zero Reconnect Time in the last Control Connection initiation message from this peer), the LCCE retains the remote information for the sessions associated with this Control Connection (rather than discarding the information), but all these sessions transit to the stale state. The LCCE SHOULD start

its reconnecting procedures immediately. Failure to reconnect MUST NOT cause termination of the Graceful Restart procedure.

The amount of time the LCCE keeps its stale sessions remote information is set to the lesser of the GR Reconnect Timeout, as was advertised by the peer, and a local timer, called the Peer Liveness Timer. If within that time the LCCE still does not establish an L2TP Control Connection with its peer, the remote information of all the stale sessions MUST be deleted and these sessions MUST transit to the idle state. The Peer Liveness Timer is started when the LCCE detects that its L2TP Control Connection with the peer went down. The value of the Peer Liveness timer SHOULD be configurable.

If the LCCE re-establishes a L2TP Control Connection with its peer within the lesser of the GR Reconnect Timeout and the Peer Liveness Timer, and the LCCE determines (by receiving Recovery Time equal to zero) that the peer was not able to preserve its L2TP forwarding state, the remote information for all the stale sessions MUST be immediately deleted and all these sessions MUST transit to the idle state. If the LCCE determines that the peer was able to preserve its L2TP forwarding state (as was indicated by the non-zero Recovery Time sent by the peer), the LCCE SHOULD further keep the stale sessions, received from the peer, for as long as the lesser of the Recovery Time advertised by the peer, and a local configurable value, called Maximum Recovery Time, allows. This value is the one set in the Recovery Time sent to the peer when re-establishing the Control Connection.

4.3 Accepting request to start Control Connection before disconnect detection

An LCCE may fully restart before its Peer LCCE detects the failure of the Control Connection. This will cause the Peer LCCE to receive a SCCRQ for a Control Connection that is still in the established state. (Handling of multiple Control Connections between a pair of LCCEs is discussed later.) If the SCCRQ contains the Graceful Restart AVP the LCCE SHOULD continue operation as described above. If the SCCRQ does not contain the Graceful Restart AVP it should handle the SCCRQ like described in [3] and tear down the control connection and all the associated sessions.

4.4 Recovering stale Sessions

After the re-establishment of Control Connection both LCCEs have marked session in stale state. From this point on re-establishment of Sessions is symmetric. For the Sessions in the stale state (stale Sessions) reconnection is similar to the normal way with the following difference:

When an LCCE sends OCRQ or ICRQ for a stale Session it MUST add the Graceful Restart Session AVP, MUST send the preserved value of Local Session ID in the Local Session ID AVP and MUST supply the preserved Remote Session ID in the Remote Session ID AVP. If the preserved Forwarding State included a cookie, its preserved value MUST be sent in the Assigned Cookie AVP instead of using a new random value. When an LCCE sends OCRQ or ICRQ for a non-stale session it MUST NOT add the Graceful Restart Session AVP and MUST follow the normal procedures for the values of Local Session ID, Remote Session ID and the cookie.

When an LCCE receives OCRQ or ICRQ with the Graceful Restart Session AVP it will search for the corresponding Session according to the value in the Remote Session ID AVP. If this value is found, the Session is in stale state and the Local Session ID value also matches then the Session is associated with the new control connection, transits to the established state and the preserved the Local session ID, Remote Session ID and the cookie are included in the corresponding AVPs. If the Session was not in the stale state or there was a mismatch in the Local Session ID value or the cookie value, the LCCE MUST tear down the Session with the CDN Message using the Result Code 2 and the Session Graceful Restart Mismatch Error Code, delete the Session remote information and put the Session in the idle state. The LCCE MAY compare other AVP values that arrive with the OCRQ or ICRQ to validate the Graceful Restart of the Session.

When LCCE receives OCRQ or ICRQ without the Graceful Restart Session AVP it will treat it as described in [3] unless the Session is in the stale state. In that case the LCCE MUST tear down the session with CDN Message using the Result Code 2 and the Session Graceful Restart Mismatch Error Code, delete the Session remote information and transit to the idle state.

4.5 Partial Graceful Restart

An LCCE MAY support partial Graceful Restart. By this we mean that it cannot preserve its own state across its own restart but it can preserve it across its peer restart. An LCCE that supports partial Graceful Restart indicates it by including the GR AVP with Reconnect Time set to zero.

4.6 Multiple Control Connections between a pair of LCCEs

L2TP supports multiple Control Connections between a given pair of LCCEs (identified by their respective Router IDs). It is up to the LCCEs to be able to associate the correct end points of each Control Connection. This can be done according to different criteria such as

Application ID AVP, Capability List AVP etc. E.g., the LCCE MAY decide that it does not allow more than one Control Connection to its peer with the same Application ID and an overlap in the Capabilities' list. The same criteria MUST be applied when restarting the Control Connection. The LCCE MUST NOT use the Control Connection ID to identify the Control Connection across restart.

5. Security Considerations

The mechanism described in this document does not add any new security considerations for L2TPv3. In particular:

o All the checks required during a regular restart are performed between the restarting LCCE and its peer in the case of Graceful Restart

o It is impossible to change any of the L2TPv3 forwarding state including source and destination IP address, Session ID and cookie values etc.

The security considerations pertaining to the original L2TP protocol [3] remain relevant.

6. IANA Considerations

This document requires assignment of the following numbers by IANA:

o Two new AVP types (see Sections 2.1 and 2.2 above)

o One new Error Code value (see Section 2.4 above).

Copyright notice

References

1 Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

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

3 J. Lau, M. Townsley, I. Goyret , ôLayer Two Tunneling Protocol (Version 3)ö, Work in Progress, draft-ietf-l2tpext-l2tp-base-14.txt, June 2004.

4 Leelanivas, M., Rekhter, Y. and R. Aggrawal, "Graceful Restart Mechanism for Label Distribution Protocol", RFC 3478, February 2003.

5 Farrel, A., "Fault Tolerance for the Label Distribution Protocol (LDP)", RFC 3479, February 2003.

6 S. Bryant, P. Pate, PWE3 Architecture, Work in Progress, draft-ietf-pwe3-arch-07.txt, March 2003

7 L. Martini et al, Pseudowire Setup and Maintenance Using LDP, Work in progress, draft-ietf-pwe3-control-protocol-11.txt, October 2004

Acknowledgments

I would like to thank Sam Henderson and Sasha Vainshtein for their constructive comments on this memo. I would like to also thank Gonen Zilber for participating in the writing of the first draft of this memo.

AuthorsÆ Addresses

Sharon Galtzur
Axerra Networks
24 Raoul Wallenberg
Tel Aviv, Israel
Email: Sharon@Axerra.com