Network Working Group
INTERNET DRAFT
Expires May 2002

Reinaldo Penno
Nortel Networks
Keyur Parikh
Megisto Systems
Ly Loi
Tahoe Networks
Leo Huber
Extreme Networks
Vipin Jain
Pipal Systems
Mark Townsley
Cisco Systems
February 2002

Fail Over extensions for L2TP
draft-vipin-l2tpext-failover-02.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC 2026. 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.

Abstract

L2TP is a connection-oriented protocol that has shared state between active endpoints. Some of this shared state is vital for operation but may be rather volatile in nature, such as packet sequence numbers used on the L2TP Control Connection. When failure of one side of a control connection occurs, a new control connection is created and associated with the old connection by exchanging information about the old connection. Such a mechanism is not intended as a replacement for an active fail over with some mirrored connection states, but as an aid just for those parameters that are particularly difficult to have immediately available. Protocol extensions to L2TP defined in this document are intended to facilitate state recovery, providing

additional resiliency in an L2TP network and improving a remote system's layer 2 connectivity.

Table of Contents

Terminology

Endpoint: An L2TP control connection endpoint, either LAC or LNS.

Active Endpoint: An endpoint that is currently providing service.

Backup Endpoint: A redundant endpoint standing by for the active endpoint.

Failover: The action of a Backup Endpoint taking over the service of an Active Endpoint. This could be due to administrative action or failure of the Active Endpoint.

1.0 Introduction

The goal of this draft is to aid the overall resiliency of an L2TP endpoint by introducing extensions to RFC 2661 [L2TP] that will minimize the recovery time of the L2TP layer after a failover, while minimizing the impact on its performance. Therefore it is assumed that the endpoint's overall architecture is also supportive in the resiliency effort.

To ensure proper operation of a L2TP endpoint after a failover, the associated information of the tunnels and sessions between them must be correct and consistent. This includes both the configured and dynamic information. The configured information is assumed to be correct and consistent after a failover, otherwise the tunnels and sessions would not have been setup in the first place. The dynamic information, which is also referred to as stateful information, changes with the processing of tunnel's control and data packets. Currently, the only such information that is essential to the tunnel's operation is its sequence numbers. For the tunnel control channel, the inconsistencies in its sequence numbers can result in the termination of the entire tunnel. For tunnel sessions, the inconsistency in its sequence, when used, can cause massive data loss thus giving perception of "service loss" to the user.

Thus, an optimal resilient architecture that aims to minimize "service loss" after a failover must make provision for the tunnel's essential stateful information - i.e. its sequence numbers. Currently, there are two options available: the first option is to ensure that the backup endpoint is in complete sync with the active with respect to the control and data sessions sequence numbers. The other option is to simply re-establish all the tunnels and its sessions after a failover. The drawback of the first option is that it adds significant performance and complexity impact to the endpoint's architecture, especially as tunnel and session aggregation increases. The drawback of the second option is that it increases the "service loss" time, especially as the architecture scales.

To alleviate the above-mentioned drawbacks of the current options, this draft introduces a mechanism to bring the dynamic stateful information of a tunnel to correct and consistent state after a failure. Proposed mechanism, currently, defines the recovery of tunnels and sessions that were in established state prior to the failure.

2.0 Failover Protocol

The failover protocol allows an endpoint to specify its failover capabilities during tunnel establishment. Based on failover capabilities, two endpoints learn if a tunnel and its sessions support recovery. Upon failure, a new tunnel is initiated for every old tunnel that needs recovery. The new tunnel includes a new AVP, the Old Tunnel ID AVP (Section 3.3). This AVP identifies the old tunnel. Upon getting this AVP, an endpoint learns that its peer has failed and would like to recover the identified tunnel. After the new tunnel is established, it assumes all active sessions and tunnel characteristics of the previous tunnel. Normal tunnel activity is resumed then.

2.1 Tunnel Establishment

Tunnel establishment procedures are same as defined by [L2TP], except when establishing a tunnel, endpoints exchange their failover capabilities using Failover Capability Initiate AVP and Failover Capability Response AVP in SCCRQ and SCCRP control messages.

2.2 Session Establishment

There is no change to how [L2TP] describes session establishment and termination procedure.

2.3 Post Failure Operation

This section describes the behavior of a failed endpoint and its peer, should there be failure. Peers must avoid sending any control packets over the old tunnel that is being synchronized. Only those tunnels, which exchanged failover capabilities are allowed to use failover protocol.

2.3.1. Failed Endpoint's Behavior

It establishes a new tunnel as specified in [L2TP] with following considerations:

- SCCRQ and SCCCN messages SHOULD avoid using new AVPs that might impact the establishment of the new tunnel or change its characteristics.
- In addition to the AVPs present in the old tunnel, it MUST include the Old Tunnel Id AVP and Old Local Tunnel Id AVP, as defined in section 3.0, in the new SCCRQ.
- If the new tunnel is rejected then it MUST assume that recovery has failed and should clear the tunnel on its end.
- Once failover is detected, control packets received prior to synchronization of the tunnel SHOULD be ignored. The tunnel is said to be in synchronized state once the establishment process is complete.

2.3.2. Failed Endpoint's Peer's behavior

It accepts tunnel requests from the peer as specified in [L2TP] with following considerations:

- It MUST use the Old Tunnel Id AVP, as described in section 3.3, to determine whether peer is trying to recover an old tunnel and if present which tunnel.
- It MUST validate Old Tunnel Id and Old Local Tunnel Id in an incoming SCCRQ. If it doesn't find a match it MUST reject the SCCRQ.
- It may choose to reject the new tunnel request if it did not advertise any failover capabilities on corresponding old tunnel.
- Upon establishment of the new tunnel, it assumes the sessions belonging to old tunnel on the new tunnel and begin normal operation on the tunnel.

2.3.3 Session State Inconsistency Between Peers

The assumption in the failover mechanism is that the backup endpoint is kept in sync with the final state of the sessions when they are established or terminated. However, it is possible that a failover may occur while sessions are in transient state, in this case an endpoint may retransmit unacknowledged control messages to bring the session states to a consistent state once the tunnel is recovered. However, in some cases, this may not be possible, for example if a LAC initiates a session and fails prior to receiving its final ACK, then the session is considered established at the LNS while at the LAC it is still in transient state and therefore not backed up. In this case, the LNS may use other means to detect the "dangling" session and locally terminate it.

2.3.4. Data Plane Behavior

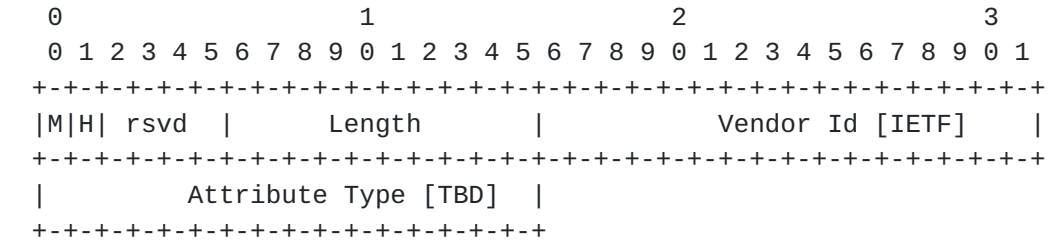
If sequencing was used on data sessions, upon detecting peer's failure, the non-failed endpoint MUST set the next expected Ns based on the incoming Ns value. It must also flush re-ordering buffers if applicable.

3.0. Failover AVPs

The new AVPs that should be included in SCCRQ, SCCRP messages are as follows:

3.1. Failover Initiate Capability AVP [SCCRQ, SCCRP]

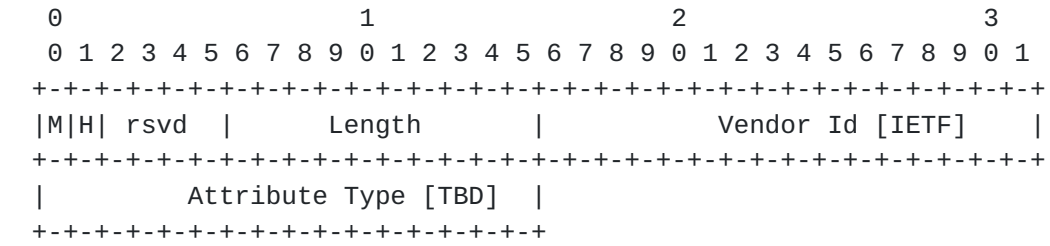
Failover Capability Initiate AVP, Attribute Type [TBD], describes if an endpoint could initiate recovery on a given tunnel after failure.



The AVP is not mandatory (the M-bit MUST be set to 0) The AVP MAY be hidden (the H-bit set to 0 or 1).

3.2. Failover Response Capability AVP [SCCRQ, SCCRP]

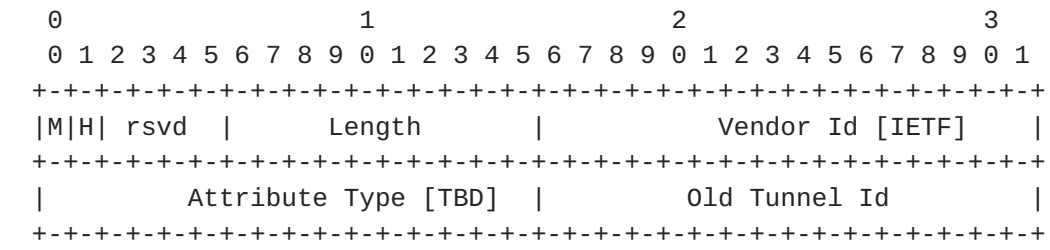
Failover Capability Response AVP, Attribute Type [TBD], describes if an endpoint is capable of responding to failure on a given tunnel.



The AVP is not mandatory (the M-bit MUST be set to 0) The AVP MAY be hidden (the H-bit set to 0 or 1).

3.3. Old Tunnel ID AVP [SCCRQ, SCCRP]

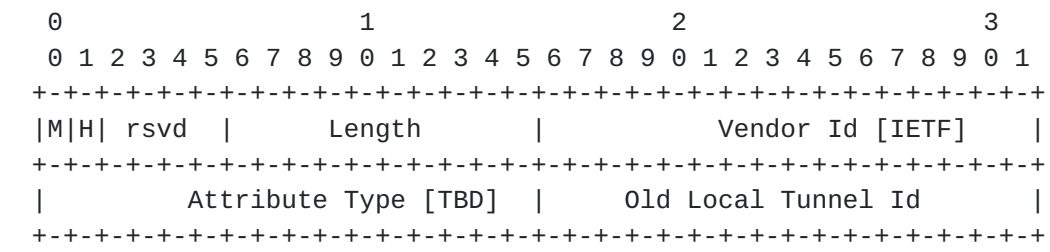
The Old Tunnel ID AVP, Attribute Type [TBD], encodes the Tunnel ID in SCCRQ and SCCRP messages that was assigned by the receiver before failure.



This AVP is mandatory(the M-bit MUST be set to 1). The AVP may be hidden (the H-bit set to 0 or 1).

3.4. Old Local Tunnel ID AVP [SCCRQ, SCCRP]

The Old Tunnel Local ID AVP, Attribute Type [TBD], encodes the Tunnel Id in SCCRQ and SCCRP messages that was assigned by the sender before failure.



The AVP is not mandatory (the M-bit MUST be set to 1) The AVP may be hidden (the H-bit set to 0 or 1).

4. IANA Considerations

This document requires three new "AVP Attributes" to be assigned through IETF Consensus [RFC2434] as indicated in Section 10.1 of [RFC2661]. These are:

Failover Initiate Capability AVP (section 3.1)

Failover Response Capability AVP (section 3.2)

Old Tunnel ID AVP (section 3.3)

Old Local Tunnel ID AVP (section 3.4)

This document defines no additional number spaces for IANA to manage.

5. Security considerations

The failover mechanism described here leaves a small (1 in 2^32) room for an intruder to discover the old tunnel id of an existing tunnel by trying out various possibilities in Old Tunnel Id and Old Local Tunnel Id AVP.

6. References

[L2TP] Townsley, et. al., "Layer Two Tunneling Protocol L2TP", RFC2661

7. Authors' Addresses

Reinaldo Penno
Nortel Networks
2305 Mission College Blvd
Santa Clara, CA 95054
Phone: +1 408.565.3023
Email: rpenno@nortelnetworks.com

Keyur Parikh
Megisto Systems
20251 Century Boulevard, Suite 120
Germantown, MD 20876
Phone: +1 301.444.1723
Email: kparikh@megisto.com

Leo Huber
Extreme Networks
3585 Monroe St.
Santa Clara CA 95051
Phone: +1 408.597.3037
Email: lhuber@extremenetworks.com

Ly Loi
Tahoe Networks
3052 Orchard Drive
San Jose, CA 95134
Phone: +1 408.944.8630
Email: lll@tahoenetworks.com

Vipin Jain
Pipal Systems
2903 Bunker Hill Ln #210
Santa Clara, CA 95054
Phone: +1 408.470.9700
Email: vipinietf@yahoo.com

W. Mark Townsley
Cisco Systems
7025 Kit Creek Road
PO Box 14987
Research Triangle Park, NC 27709
EMail: townsley@cisco.com