Network Working Group
Internet Draft
Expiration Date: February 2004

Eric C. Rosen
Cisco Systems, Inc.

August 2003

Detecting and Reacting to Failures of the Full Mesh in IPLS and VPLS

draft-rosen-l2vpn-mesh-failure-00.txt

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.

Abstract

Certain L2VPN architectures [IPLS, VPLS] rely on there being a full mesh of pseudowires [PWE3-ARCH] among a set of entities. This mesh is used to provide a "LAN-like" service among the entities. If one or more of these pseudowires is absent, so that there is not really a full mesh, various higher layers (from routing to bridge control protocols) that expect a LAN-like service may fail to work as expected. Therefore it is desirable to have procedures that enable the pseudowire endpoints to determine automatically whether there is really a full mesh or not. It is also desirable to have procedures that cause the L2VPNs to adapt to pseudowire failures. This document proposes a set of procedures to meet these goals. Detailed protocol encodings are not present, but will be added in future versions.

Contents

1 Introduction 2
2 Detection of Partially Connected EEs 4
3 Actions Taken Upon Detection 5
4 References 7
5 Author's Information 7

1. Introduction

IPLS [IPLS] interconnects a set of CEs. With respect to a particular IPLS instance and a particular PE supporting that IPLS instance, the set of CEs can be divided into the PE's "local CEs" and the PE's "remote CEs". The local CEs are directly attached to the PE. ("Directly attached" means attached via an "Attachment Circuit" in the sense of [L2VPN-Framework].) The PE must ensure that each of its local CEs is bound, by a Pseudowire (PW), to each of the remote CEs. When this condition holds for all the PEs supporting a given IPLS instance, we say that the IPLS instance is fully meshed.

VPLS [VPLS} interconnects a set of "VPLS Forwarders" [L2VPN-FRAMEWORK], which are virtual entities inside PEs; for a given VPLS instance, there is one VPLS Forwarder in a given PE. Some of these are considered "spokes", and some are considered "hubs". In a given VPLS instance, there must be a PW binding every hub VPLS Forwarder to every other hub VPLS Forwarder; this means that every hub PE in the VPLS instance must have a PW to every other hub PE in the VPLS instance. When this condition holds, we say that the VPLS instance is fully meshed.

We will use the term "LS" to mean "IPLS or VPLS".

In each LS instance, there is a set of "endpoint entities" (EEs). In VPLS, the EEs are hub VPLS Forwarders inside the PEs, in IPLS the EEs are CEs. In either case, we say say that the LS instance is "fully meshed" if every pair of EEs which are not local to the same PE are bound together by a PW.

(For present purposes, it does not matter whether two EEs are bound by a single bidirectional point-to-point PW or by a pair of unidirectional point-to-multipoint PWs.)

It is possible that a given LS instance may fail to be fully meshed. This may happen for the following reasons:

- Configuration errors.

- Failure of the auto-discovery process.

- Failure of the control plane to properly establish all the necessary PWs. This in turn may be due to bugs, or to resource shortages at the PEs.

- Failure of the data plane to carry traffic correctly on all the established PWs. This can occur if there are bugs in the encapsulation/decapsulation procedures at the PEs, or bugs in the forwarding procedures at intermediate nodes (especially in technologies where the data and control planes are decoupled.

When an LS instance is not fully meshed, we will say that one or more of its EEs are "partially connected". An EE is regarded as "partially connected" at a particular time if one of the following conditions holds:

- PW not established: at that time, some PW binding that EE to another EE has not been properly established, as determined by the PW control plane.

- PW not operational: at that time, although the control plane indicates that all the PWs binding other EEs to the given EE are properly established, one or more of those PW is incapable of passing data to the given EE for some reason. Note that "operational" status is a unidirectional attribute.

If an LS instance is not fully meshed, then it will not be able to provide the "LAN-like" service on which its users are depending. For instance, if a link state routing algorithm is using its LAN procedures over an LS instance which is not fully meshed, the selected set of routes may have "black holes".

It is desirable therefore to have procedures which will automatically identify any partially connected EEs. This document proposes a set of procedures to meet these goals. Detailed protocol encodings are not present, but will be added in future versions if the WG has interest in proceeding in this direction.

2. Detection of Partially Connected EEs

Each PE in a particular LS instance must have some sort of control plane relationship with each of the other PEs in the same LS instance. (For the time being we ignore the situation in which PWs are spliced together; this concepts discussed here are readily extended to that case.)

There must be a status message, which we call the "Mesh Status" message, which a PE sends to each of the other PEs in the same LS instance. The Mesh Status message identifies the LS instance (by its globally unique VPN identifier, for example), and lists the set of EE pairs for which the originating PE has operational PWs. This message would need to be resent whenever the list changes. As long as the control protocol can reliably transport control messages, this message would not have to be sent unless there is a change; in fact, only changes would need to be sent. (However, this would require two variants of the Mesh Status message: an "Add" and a "Remove".) A PE's Mesh Status messages should also indicate which of the EEs are locally attached to that PE.

Thus every PE in an LS instance maintains the Mesh Status of every other PE supporting that same LS instance.

When the control connection to a particular remote PE is lost, the Mesh Status of the remote PE is flushed, and no longer considered for the purposes of Partially Connected EE Detection.

By including a pair of EEs in its Mesh Status messages, a PE is stating that there is an OPERATIONAL PW binding the two EEs together, not merely an established PW. Each PE is responsible for determining whether each of its local PWs is operational in the outgoing direction. This may require the use of some sort of per-PW test of the data plane. It is advisable to construct the test for operational status so as to avoid the possibility of flapping, perhaps by not allowing a non-operational PW to return to operational status in less than a specified time period. The test for operational status should also ensure that a PW is not declared non-operational due to ordinary network conditions, such as occasional packet loss, and that a PW is not declared non-operational due to routing transients.

It is understood that it is much easier to lay down such requirements than it is to devise procedures to meet them. The specification of such procedures however is outside the scope of the current document.

When a PE in a particular LS instance has received a Mesh Status message from every other PE (that it knows about) in that instance, it can compute the set {EE} of all the EEs in the LS instance. This

is the union of the set of EEs mentioned in all the Mesh Status messages.

The IPLS or VPLS instance is fully meshed if and only if the following condition holds:

> For every PE p and every EE e, either e is one of p's local EEs, or p reports an operational PW from each of its local EEs to e.

If this condition doesn't hold, there are one or more Partially Connected PWs . The set of Partially Connected EEs is defined as follows:

> An EE e is "Partially Connected" if and only if there is some PE p such that e is not locally attached to p, and p has a locally attached EE e' such that there is either no operational PW from e to e' or there is no operational PW from e' to e.

If the configuration and/or auto-discovery procedures identify a set of EEs whose local PE just happens to be down (or otherwise unreachable), no PEs will have operational PWs for any of those EEs, and the above procedures will not result in the determination that there are any Partially Connected EEs. However, misconfigurations or auto-discovery problems which cause different PEs to learn about different sets of EEs will result in the detection of Partially Connected EEs.

3. Actions Taken Upon Detection

Upon identification of a Partially Connected EE, an alarm should be raised so that the network operators are aware of the situation.

In general, the LS service will not function properly if there are Partially Connected EEs. It can however be made to function properly if the Partially Connected EEs are removed from service entirely, until such time as they becomes fully connected. In effect, once the problematic EEs are removed from the mesh entirely, the LS service is once again fully meshed, though with fewer EEs. Any users who connect via the removed EEs will of experience degraded service, if not complete loss of service, but other users may continue to receive service.

If a PE determines that one of its locally attached EEs is Partially Connected, it should remove that EE from service. In the case of VPLS, this means that an Emulated LAN interface [L2VPN-Framework] is brought down. In the case of IPLS, this means that the Attachment Circuit to a particular set of CEs is brought down. PWs which are

bound to the Emulated LAN interface or Attachment Circuit should NOT be disestablished and the testing of the data plane of such PWs should continue.

If a PE determines that a remote EE is Partially Connected, the PE will cease to send or receive data to or from that EE. The corresponding PWs should NOT be disestablished, and the testing of the data plane of such PWs should continue.

There may be methods of returning the LS service to a full mesh which do not require removing a Partially Connected EE from service entirely. For example, in VPLS it may be possible to change a Partially Connected EE from a hub to a spoke, thereby removing it from the mesh without bringing it out of service. [HUB-TO-SPOKE]

If, at some later time, an EE ceases to be Partially Connected, normal operations can resume.

It must be understood that when an EE first becomes known, there will be a period of time during which PEs are trying to bring up PWs to it. From the time the first PW to/from it becomes operational to the time the last PW to/from it becomes operational, the EE will be detected as Partially Connected. As this is a normal transient, there should be a specified period of time during which a newly discovered EE may be Partially Connected before any action is taken. Determination that a previously known EE has become Partially Connected should cause immediate actions, however.

If a PE detects that one of its PWs has ceased to be operational, the remote EE does not necessarily get treated immediately as being Partially Connected. Before declaring the EE to be Partially Connected, the PE should wait a period of time to see if that EE disappears from the Mesh Status messages generated by all the other PEs. After all, a very likely cause for a PW to become non-operational is for the remote PE to fail or to become unreachable. As this will no result in a partial mesh, no special action needs to be take.

4. References

[HUB-TO-SPOKE] as suggested by Vach Kompella on the L2VPN mailing list

[IPLS] "IP over LAN Service (IPLS)", H. Shah, K. Arvind, E. Rosen, G. Heron, V. Radoaca, draft-shah-ppvpn-ipls-02.txt, June 2003

[L2VPN-FRAMEWORK] "L2VPN Framework", L. Andersson, E. Rosen, editors, draft-ietf-l2vpn-l2-framework-00.txt, February 2003

[PWE3-ARCH] "PWE3 Architecture", S. Bryant, P.Pate, editors, draft-ietf-pwe3-arch-04.txt, June 2003

[VPLS] "Virtual Private LAN Services over MPLS", M. Lasserre, V. Kompella, et. al., draft-ietf-l2vpn-vpls-ldp-00.txt, June 2003

5. Author's Information

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719

E-mail: erosen@cisco.com