Network Working Group
INTERNET-DRAFT
Expires January 2002
Updates RFC 1990

J. Carlson
Sun Microsystems
July 2001

PPP Link Balancing Detection (LBD)
<draft-ietf-pppext-lbd-03.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.

This document is the product of the Point-to-Point Protocol Extensions Working Group of the Internet Engineering Task Force (IETF). Comments should be submitted to the ietf-ppp@merit.edu mailing list. Distribution of this memo is unlimited.

Abstract

The Point-to-Point Protocol (PPP) [1] provides a standard method for transporting multi-protocol datagrams over point-to-point links. PPP also defines an extensible Link Control Protocol (LCP), that allows the detection of optional link handling procedures, and a Multilink procedure (MP) [2], that allows operation over multiple parallel links. This document defines an extension to MP called Link Balancing Detection (LBD) and the LCP options that control this extension. This extension allows high-speed implementations to use the single-NCP negotiation model of MP without requiring the cost associated with the unneeded MP datagram buffering and reordering.

Table of Contents

1. Introduction

PPP negotiation allows for two types of links with regard to multiple link layer entities. A non-MP PPP link is negotiated without the Maximum-Receive-Reconstructed-Unit (MRRU) option and appears as a separate network interface to the network layer and to routing protocols. The Multilink PPP (MP) [2] type of link uses the MRRU option and allows multiple PPP links to be bundled into one network interface. An MP bundle appears as a single network interface to the network layer and to routing protocols.

There are many advantages having multiple links between two nodes appear at the network layer to be a single link. While equal-cost multi-path balancing is certainly possible with modern interior gateway protocols, less stress is placed on scarce routing system resources when link-layer detection of parallel links is employed, allowing current routing protocols to scale better. Also, the routing protocols are more stable when individual link failures are not visible to link-state routing protocols.

The main disadvantage to the current MP technique is that it does not constrain the fragmentation that may be done by the peer. For systems employing general purpose CPUs in the data path and with scatter-gather direct memory access (DMA) capability, the reassembly of datagrams fragmented on arbitrary octet boundaries is often not a problem. For systems with very high bandwidth capabilities, these features are often infeasible, and this problem makes regular MP unusable.

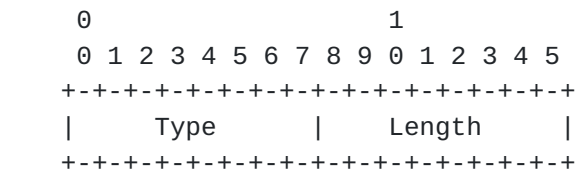
This document describes a method similar to and compatible with MP's algorithm to detect parallel links between two nodes, but without the MP headers and the associated requirement of fragmentation and sequencing. Instead, datagrams are distributed without MP headers among the links in the bundle in any convenient manner, including based on a flow-identifying hash or on round-robbin service, as long as the chosen mechanism is sufficient for the supported network layer protocols.

This technique is also referred to as "load balancing." The difference between LBD and traditional load balancing is that MP's single-NCP model (and associated single network layer address) is used, the configuration of the parallel links is made automatic, and configuration errors are detected. This allows peers to discover during LCP negotiation that, for example, links within a configured bundle violate an implementation constraint by having different MRU values, or are provisioned to terminate on the wrong network node.

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 [3].

2. No-Fragmentation Configuration Option Format

A summary of the No-Fragmentation Configuration Option format for LCP is shown below. The fields are transmitted from left to right.



Type

TBD-1

Length

2

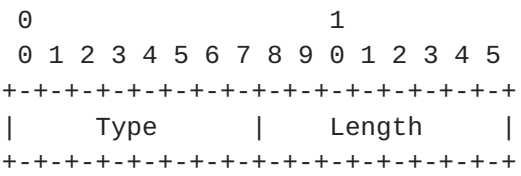
The sender of this option in an LCP Configure-Request message is indicating to its peer that it cannot support MP reassembly, and, thus the peer must not fragment messages that it sends.

If the peer Configure-Ack's this option, then the peer MUST NOT fragment frames using MP fragmentation. It MAY still use MP headers to

preserve frame sequencing. If the peer Configure-Reject's this option, then the sender must remove this option from its next Configure-Request message and MAY decline to run MP by also removing its MRRU Configuration Option. Implementations MUST NOT Configure-Nak this option if it appears in the peer's Configure-Request.

3. No-MP-Headers Configuration Option Format

A summary of the No-MP-Headers Configuration Option format for LCP is shown below. The fields are transmitted from left to right.



Type

TBD-2

Length

2

The sender of this option in an LCP Configure-Request message is indicating to its peer that it cannot support standard MP headers, and thus the peer must not use MP headers on the messages that it sends, and must send network layer messages using their assigned protocol numbers rather than inside protocol 003D.

If this option is specified, then the No-Fragmentation option is unnecessary. Fragmentation without MP headers is not supported.

If the peer Configure-Ack's this option, then it MUST NOT add MP headers or fragment frames using MP. If the peer Configure-Reject's this option, then the sender must remove this option from its next Configure-Request message and MAY decline to run MP by also removing its MRRU Configuration Option. Implementations MUST NOT Configure-Nak this option if it appears in the peer's Configure-Request.

This option MUST NOT be used on links that are intended to carry network protocols that cannot tolerate reordering, such as Bridging [4]. See section 6 of this document for details.

4. Interactions

4.1. Interaction With MRRU and MRU

The MRRU option from MP is used to signal the desire to run MP, regardless of whether or not these options are present, and to set the maximum network layer MTU. If the MRRU option is not negotiated, then MP is not enabled and the options in this document have no effect. If an MRRU is negotiated, then the MTU advertised to the local network layer by PPP MUST NOT be greater than the peer's MRRU.

If the No-MP-Headers option is used, then the MTU MUST also be limited by the peer's MRU. That is, the MTU is calculated as the minimum of the peer's MRU and MRRU.

If the No-Fragmentation option is used without the No-MP-Headers option, the MTU MUST also be limited by the peer's MRU minus the MP overhead (6 octets for the default long sequence numbers, 4 octets for the optional short sequence numbers).

4.2. Interaction With CCP and ECP

The No-Fragmentation option has no effect on either CCP or ECP. However, when the No-MP-Headers option is negotiated, reordering is possible. To avoid harmful interaction with these protocols, one of the following mechanisms MAY be used:

- Attempt to negotiate the per-link forms of CCP and ECP first.

- If bundle-level CCP or ECP is required, negotiate to use only history-less algorithms.

- If history-less algorithms are unavailable or disabled, then algorithms
  supporting multiple contexts in an implementation-dependent manner
  MAY be used by prior arrangement.

Otherwise, CCP and ECP should be disabled. Since disabling of MP headers is intended for use with high-speed links where CCP and ECP are also problematic, this last option is RECOMMENDED.

4.3. Interaction With IGPs

MP bundling has two desirable effects on link-state algorithms.

First, by summarizing the physical links into a single virtual link for IP, it reduces the size of the Router LSAs carried. Second, by leaving the IP link existence unchanged when member links join and leave the bundle, it reduces the need to advertise changes in the Router LSAs and the associated network and CPU overhead due to SPF recalculation.

It has been suggested that extensions to the OSPF and IS-IS neighbor discovery process could perform the same function as LBD, perhaps in a manner patterned after PNNI. However, doing this would require both new features in each of the link state protocols and changes in the way LSAs are constructed. The LBD mechanism is proposed as a simpler solution.

As with link-state routing protocols, MP bundling also improves the stability of the distance-vector routing protocols, such as RIPv2. LBD also adds the ability to handle parallel links, which generally cannot be used by these protocols.

4.4. Interaction With QoS

The properties of the bundle may be summarized for admittance control by advertising the aggregate bandwidth and maximum reservable bandwidth among the member links. Detailed QoS specification is outside the scope of this document.

5. Bundle Establishment and Tear-Down

As with MP, bundle establishment is based on a combination of the peer's supplied MP Endpoint Discriminator (ED) and the peer's identity as determined via link authentication. The algorithm used for LBD is identical to the MP algorithm, and is documented here only for convenience.

When authentication (if any was negotiated via LCP) is complete, a check is made before attempting to negotiate any Network Control Protocols (NCPs). If an MRRU is agreed to by both peers and if there is an existing LBD bundle where the ED (or lack thereof) matches the new link's ED (or lack), and the authenticated peer name (or lack thereof) match the new link's peer name (or lack), then this new link should be made part of the bundle and no new NCPs are created. Otherwise, this is a separate link, and NCPs should be started.

If the local and remote MRRU values do not agree with the bundle or if the presence or absence of the No-Fragmentation or No-MP-Headers options does not agree with the bundle, then the link SHOULD be

terminated. An implementation MAY choose instead to renegotiate LCP to repair the error.

Tear-down is identical to standard MP and is thus not covered here.

6. Message Distribution

To distribute messages among the links when LBD is in effect, a few simple rules must be followed.

First, since PPP negotiation does not withstand reordering, all PPP negotiation messages MUST be sent over a single link to avoid possible reordering. The first link in a bundle MUST be used to transmit PPP messages until this link is terminated. If the first link is terminated, then one remaining link in the bundle MUST be chosen for subsequent messages. Once that link is chosen, an implementation MUST continue sending all PPP negotiation messages over that single link. Any remaining link in the bundle MAY be chosen, and it is entirely possible that each peer may choose a different link without harm to PPP.

Second, PPP negotiation messages MUST be handled when received on any link. An implementation MAY choose to terminate the last link over which negotiation was received if a negotiation message is received over a different link, since this transition implies that the peer has already terminated the prior link.

Third, network datagrams SHOULD be distributed over all links as evenly as possible. There are no requirements that any particular distribution algorithm be used. Note, however, that some network protocols behave poorly when subjected to message reordering, thus techniques that reduce the likelihood of reordering (such as deterministic hashes of network layer addresses and transport identifiers) are encouraged. For TCP, reordering of IP datagrams usually causes a "slow path" in most implementations to be taken, and can trigger undesirable side-effects, such as fast retransmit.

Fourth, network datagrams from protocols that cannot withstand message reordering MUST be sent over a single link within the bundle. The link for each datagram may be chosen in any manner appropriate for that network layer, and is left to either the network layer specification or prior arrangement between the peers. For instance, an implementation using bridging with VLANs could allocate the VLANs among the available links using the same algorithm as described for PPP negotiation messages. Avoiding the use of the No-MP-Headers option may be preferred in these cases, since the changes to the hash-to-link mapping required when links join or leave the bundle can

cause small amounts of reordering.

Fifth, the common but technically non-standard practice of using LCP Terminate-Request to terminate a link gracefully without data loss is encouraged in LBD implementations. To do this, an implementation leaves Open state on sending LCP Terminate-Request, but, contrary to RFC 1661, continues processing received datagrams until the peer replies with LCP Terminate-Ack.

7. Prior Art

The traditional way to implement load balancing for IP over PPP in the absence of LBD is to allow the multiple PPP links to negotiate the same pair of IP endpoint addresses independently. The traditional mechanism has the advantage of simplicity, as it requires no protocol changes, but does not necessarily work when non-IP network protocols are in use, and does not have the routing features of LBD. When that style of load balancing is used, the member links may be advertised as separate unnumbered links.

An implementation that refuses to use the IPCP IP-Address option or uses only MPLS cannot use the traditional method and must rely on manual configuration or LBD.

8. Security Issues

The authentication and bundling techniques are identical to standard MP and the security issues are the same as with RFC 1990.

9. Acknowledgements

The idea of link-detected balancing itself was inspired by Ross Callon. I am also grateful for the many critiques and ideas offered on the IETF PPP Extensions mailing list and by private mail. In particular, I thank John Bray, Archie Cobbs, and Vernon Schryver.

10. References

[1] W. Simpson, "The Point-to-Point Protocol (PPP)", RFC 1661, 07/21/1994

[2] K. Sklower, et al., "The PPP Multilink Protocol (MP)", RFC 1990, 08/1996

[3] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels," BCP 14 and RFC 2119, 03/1997

[4] M. Higashiyama and F. Baker, "PPP Bridging Control Protocol (BCP)," RFC 2878, 07/2000.

11. Author's Address

James Carlson
Sun Microsystems
1 Network Drive MS UBUR02-212
Burlington MA 01803-2757

Phone: +1 781 442 2084
Fax: +1 781 442 1677
Email: james.d.carlson@sun.com