Network Working Group
Request for Comments: 5640
Category: Standards Track

C. Filsfils
P. Mohapatra
C. Pignataro
Cisco Systems

August 2009

Load-Balancing for Mesh Softwires

Abstract

Payloads transported over a Softwire mesh service (as defined by BGP Encapsulation Subsequent Address Family Identifier (SAFI) information exchange) often carry a number of identifiable, distinct flows. It can, in some circumstances, be desirable to distribute these flows over the equal cost multiple paths (ECMPs) that exist in the packet switched network. Currently, the payload of a packet entering the Softwire can only be interpreted by the ingress and egress routers. Thus, the load-balancing decision of a core router is only based on the encapsulating header, presenting much less entropy than available in the payload or the encapsulated header since the Softwire encapsulation acts in a tunneling fashion. This document describes a method for achieving comparable load-balancing efficiency in a network carrying Softwire mesh service over Layer Two Tunneling Protocol - Version 3 (L2TPv3) over IP or Generic Routing Encapsulation (GRE) encapsulation to what would be achieved without such encapsulation.

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Table of Contents

1. Introduction

Consider the case of a router R1 that encapsulates a packet P into a Softwire bound to router R3. R2 is a router on the shortest path from R1 to R3. R2’s shortest path to R3 involves equal cost multiple paths (ECMPs). The goal is for R2 to be able to choose which path to use on the basis of the full entropy of packet P.

This is achieved by carrying in the encapsulation header a signature of the inner header, hence enhancing the entropy of the flows as seen by the core routers. The signature is carried as part of one of the fields of the encapsulation header. To aid with better description in the document, we define the generic term "load-balancing field" to mean such a value that is specific to an encapsulation type. For example, for L2TPv3-over-IP [RFC3931] encapsulation, the load-balancing field is the Session Identifier (Session ID). For GRE [RFC2784] encapsulation, the Key field [RFC2890], if present, represents the load-balancing field. This mechanism assumes that core routers base their load-balancing decisions on a flow definition that includes the load-balancing field. This is an obvious and generic functionality as, for example, for L2TPv3-over-IP tunnels, the Session ID is at the same well-known constant offset as the TCP/UDP ports in the encapsulating header.

The Encapsulation SAFI [RFC5512] is extended such that a contiguous block of the load-balancing field is bound to the Softwire advertised by a BGP next-hop. On a per-inner-flow basis, the ingress Provider Edge (PE) selects one value of the load-balancing field from the block to preserve per-flow ordering and, at the same time, to enhance the entropy across flows.

1.1. Requirements Language

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 RFC 2119 [RFC2119].

2. Load-Balancing Block sub-TLV

This document defines a new sub-TLV for use with the Tunnel Encapsulation Attribute defined in [RFC5512]. The new sub-TLV is referred to as the "Load-Balancing Block sub-TLV" and MAY be included in any Encapsulation SAFI UPDATE message where load-balancing is desired.

The sub-TLV type of the Load-Balancing Block sub-TLV is 5. The sub-TLV length is 2 octets. The value represents the length of the block in bits and MUST NOT exceed the size of the load-balancing field. This format is very similar to the variable-length subnet masking (VLSM) used in IP addresses to allow arbitrary length prefixes. The block is determined by extracting the initial sequence of ’block size’ bits from the load-balancing field.

If a load-balancing field is not signaled (e.g., if the encapsulation sub-TLV is not included in an advertisement as in the case of GRE without a Key), then the Load-Balancing Block sub-TLV MUST NOT be included.

The smaller the value field of the Load-Balancing Block sub-TLV, the larger the space for per-flow identification, and hence the better entropy for potential load-balancing in the core, as well as, the lower the polarization when mapping flows to ECMP paths. However, reducing the load-balancing block size consumes more L2TPv3 Session IDs or GRE Keys, resulting in potentially less numbers of supported services. A typical deployment would need to arbitrate between this trade-off.

As an example, assume that there is a Softwire set up between R1 and R3 with L2TPv3-over-IP tunnel type. Assume that R3 encodes the Session ID with value 0x1234ABCD in the encapsulation sub-TLV. It also includes the Load-Balancing Block sub-TLV and encodes the value 24. This should be interpreted as follows:

- If an ingress router does not understand the Load-Balancing Block sub-TLV, it continues to use the Session ID 0x1234ABCD and encapsulates all packets with that Session ID.

- If an ingress router understands the Load-Balancing Block sub-TLV, it picks the first 24 bits out of the Session ID (0x1234AB) to be used as the block and fills in the lower-order 8 bits with a per-flow identifier (e.g., it can be determined based on the inner packet’s source, destination addresses, and TCP/UDP ports). This selection preserves the per-flow ordering of packets.

This requirement and solution applies equally to GRE where the Key plays the same role as the Session ID in L2TPv3.

Needless to say, if an egress router does not support the Load-Balancing Block sub-TLV, the Softwire continues to operate with a single load-balancing field with which all ingress routers encapsulate.

2.1. Applicability to Tunnel Types

The Load-Balancing Block sub-TLV is applicable to tunnel types that define a load-balancing field. This document defines load-balancing fields for tunnel types 1 (L2TPv3 over IP) and 2 (GRE) as follows:

- L2TPv3 over IP - Session ID. Special care needs to be taken to always create a non-zero Session ID. When an egress router includes a Load-Balancing Block sub-TLV, it MUST encode the Session ID field of the encapsulation sub-TLV in a way that ensures that the most significant bits of the Session ID, after extracting the block, are non-zero.

- GRE - GRE Key

This document does not define a load-balancing field for the IP-in-IP tunnel type (tunnel types 7). Future tunnel types that desire to use the Load-Balancing Block sub-TLV MUST define a load-balancing field that is part of the encapsulating header.

2.2. Encapsulation Considerations

Fields included in the encapsulation header besides the load-balancing field are not affected by the Load-Balancing Block sub-TLV. All other encapsulation fields are shared between variations of the load-balancing field. For example, for the L2TPv3-over-IP tunnel type, if the optional cookie is included in the encapsulation sub-TLV by the egress router during Softwire signaling, it applies to all the "Session ID" values derived at the ingress router after applying the load-balancing block as described in this document.

3. IANA Considerations

IANA has assigned the value 5 for the Load-Balancing Block sub-TLV, in the BGP Tunnel Encapsulation Attribute Sub-TLVs registry (number space created as part of the publication of [RFC5512]):

Sub-TLV name	Value
Load-Balancing Block	5

4. Security Considerations

This document defines a new sub-TLV for the BGP Tunnel Encapsulation Attribute. Security considerations for the BGP Encapsulation SAFI and the BGP Tunnel Encapsulation Attribute are covered in [RFC5512]. There are no additional security risks introduced by this design.

5. Acknowledgements

The authors would like to thank Stewart Bryant, Mark Townsley, Rajiv Asati, Kireeti Kompella, and Robert Raszuk for their review and comments.

6. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, September 2000.

[RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

Authors' Addresses

Clarence Filsfils
Cisco Systems
Brussels,
Belgium

EMail: cfilsfil@cisco.com

Pradosh Mohapatra
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

EMail: pmohapat@cisco.com

Carlos Pignataro
Cisco Systems
7200 Kit Creek Road, PO Box 14987
Research Triangle Park, NC 27709
USA

EMail: cpignata@cisco.com