Network Working Group
Internet-Draft
Intended status: Proposed Standard
Expires: November 2, 2016

A. Verma
Juniper Networks

J. Drake
Juniper Networks

R. Molina
Ericsson Inc.

W. Lin
Juniper Networks

May 2, 2016

Vpls Best-site id
draft-anshuverma-bess-vpls-best-site-id-02.txt

Abstract

With network-based applications becoming prevalent, solutions that provide connectivity over wide area become more attractive for customers. In small-to-medium enterprise sector, Virtual Private LAN Service (VPLS), is a very useful service provider offering. It creates an emulated LAN segments fully capable of learning and forwarding Ethernet MAC addresses.

Today, in VPLS implementations, within the context of a VPLS PE (VE), a single-site is selected from which all PWs are rooted. The site-election mechanism is usually hard-coded by different vendors (e.g. minimum or maximum site-id), and as such, is outside end-users control. This offers no flexibility to end-users as it forces them to define the site-id allocation scheme well in advance, or deal with the consequences of a suboptimal site-id election. Moreover, whenever the elected site-id is declared down, the traffic to and from all other sites hosted within the same VE is impacted as well.

This draft defines protocol extensions to keep core-facing pseudowires (PWs) established at all times, regardless of the events

taking place on the attachment-circuit (AC) segment when using the BGP-based signaling procedures.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

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 Internet-Draft will expire on November 2, 2016.

Copyright Notice

Table of Contents

1. Introduction

As the popularity of VPLS services continue to expand, Service Provider requirements for a scalable multi-homed solution are becoming increasingly demanding. As dictated by RFC4762 BGP-VPLS RFC, every PE participating in a VPLS domain must be fully meshed through a bidirectional pseudowire (PW). This set of PWs is built attending to the signaling information (label-block) advertised by each PE. The label-block used to build any given PW, will be the one matching the local site being elected as ’representative’ of the VPLS domain within a given PE. As stated in RFC4762, if this site is ever declared ’down’, a compliant implementation will need to either withdraw the corresponding label-block, or announce that the affected site is no longer reachable. In either case, the PW will end up being destroyed, which will have a considerable impact on other local sites relying on this specific PW. Furthermore, as a considerable amount of cycles are spent in destroying/re-building affected PWs, the overall convergence period will be severely impacted for those critical multi-homed sites that need a rapid transition to a backup PE.

This draft defines protocol extensions to keep core-facing pseudowires established at all times, regardless of the events taking place on the attachment-circuit segment when using the BGP-based signaling procedures defined in [RFC4761].

Today, in VPLS implementations, within the context of a VPLS_PE (VE), a single-site is selected from which all PWs are rooted. The site-election mechanism is usually hard-coded by different vendors (e.g. minimum or maximum site-id), and as such, is outside end-users control. This offers no flexibility to end-users as it forces them to define the site-id allocation scheme well in advance, or deal with the consequences of a suboptimal site-id election. Moreover, whenever the elected site-id is declared down, the traffic to and from all other sites hosted within the same VE is impacted as well.

In BGP VPLS MH scenarios the above pitfalls are specially acute, as not only we need to factor in the cost to bring the active PW down and run DF election in primary PE, but also in the n-DF PE and all remote-PEs within the VPLS domain. Taking into account that control-plane operation is signaled through BGP protocol, is fare to expect that many of these operations will be carried out in sequence and not in parallel, so the overall cost is usually pretty considerable in scaling scenarios.

To achieve minimal traffic disruption, this draft introduces a virtual or dummy site which will serve as the preferable or best site within each VE. Thereby, its corresponding site-id value will be defined by the end-user. But more than providing greater provisioning flexibility, the real advantage of this best-site solution relies on the capability to maintain VPLS PWs established at all times regardless of the fluctuations in AC segments.

To summarize, this best-site feature offers:

* Greater provisioning flexibility.
* Minimal traffic disruption for non-preferable sites in multi-site VEs (upon AC going down).
* Convergence period would be considerably reduced in MH setups during transient intervals.

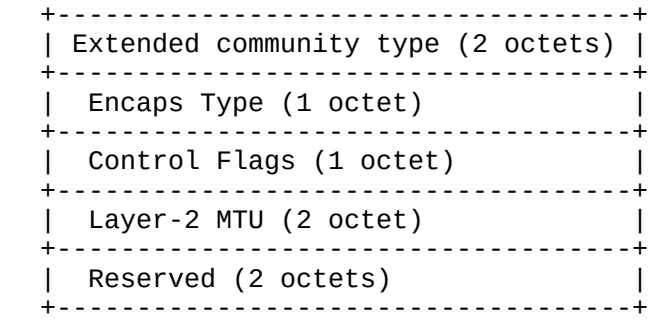
2. Conventions used in this document

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. Modifications to Layer 2 Info Extended Community

The Layer 2 Info Extended Community is used to signal control information of the pseudowires to be setup. The extended community format is described in [RFC4761]. This draft recommends that the Control Flags field of this extended community be used to synchronize the best-site information amongst PEs for a given L2VPN.



Layer-2 Info Extended Community:

Control Flags Bit Vector:

This field contains bit flags relating to pseudowire’s control information. It is augmented with the definition of one new flag field. If on a given PE VPLS instance is configured with ’best-site’, it will include in its VPLS BGP NLRI a Layer 2 Info Extended Community using Control Flags field with B = 1.

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|D|A|F|B|T|R|C|S| (Z = MUST Be Zero)
+-+-+-+-+-+-+-+-+
```

With reference to the Control Flags Bit Vector, the following bits in the Control Flags are defined; the remaining bits, MUST be set to zero when sending and MUST be ignored when receiving this Extended Community. The signaling procedure described here is therefore backwards compatible with existing implementations.

D Defined in l2vpn-vpls-multihoming draft

A Defined in l2vpn-auto-site-id draft

F Defined in l2vpn-vpls-multihoming draft

B When the bit value is 1, the PE receiving the label-block will deem the corresponding site as the most preferable site from the remote neighbor.
When the bit value is 0, the PE receiving the label-block will rely on its legacy/default site-election algorithm.

T/R Defined in l2vpn-fat-pw-bgp draft

C Defined in [RFC4761]

S Defined in [RFC4761]

4. Best-site functionality:

Traditionally, vpls path selection mechanism pick the minimum (or maximum) site-id to determine the ’preferable’ local site. This ’preferable’ local site serves two purposes: 1) pseudowires created from the local VE will be rooted from this site, and 2) pseudowires created from remote VEs will be built towards this elected site.

In order to provide some greater flexibility in the current pre-defined site-election process, this draft proposes a solution to give priority to these ’best-sites’ in detriment of those local sites with minimum (maximum) site-ids.

This solution would be fully backward compatible as VPLS-PEs on which the proposed feature isn’t enable, would simply obviate the BGP extensions previously described, and thereby, would rely on their legacy/default site-election mechanism.

Let’s make use of the following example to describe our solution in more details:

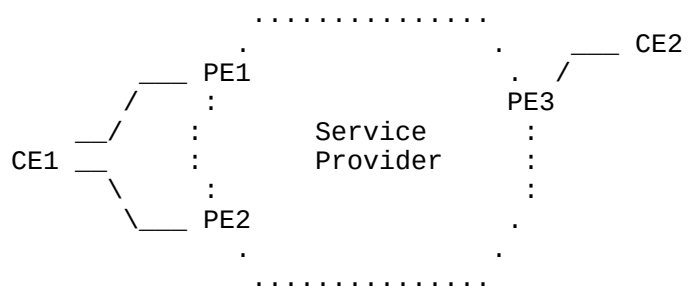


Figure 1- MH scenario with Best-site capable nodes.

A PE where ’best-site’ feature is enabled in VPLS instance, behaves as a dummy site and no access interface will be associated with it. This dummy site won’t be subjected to access interface down/up events; thereby, the corresponding D-bit will not be set to represent a site-down condition. The main goal here is to have a site that is permanently alive, regardless of the state of the attached circuits defined within the VPLS domain.

Each VPLS instance where a ’best-site’ is defined (e.g. PE1), will signal the site’s existence by setting the B-bit of the control-flags bit-vector within the L2-info extended community. Upon arrival of this BGP advertisement to the receiving PE (e.g. PE3), and only if this one is ’best-site’ capable, the received B-bit will be honored and the corresponding site will be elected as the most preferable site within the remote VE (PE1).

For those neighbors where 'best-site' feature is not configured, conventional local site election will take place. For instance, if PE1 does not receive a Label-Block advertisement with B-bit set from a remote PE (PE3), it will assume that PE3 is not 'best-site' capable, and will create a pseudowire from its minimum (maximum) designated site. For the rest of the 'best-site' capable PEs, PE1 will construct pseudowires rooted at its 'best-site' site.

By proceeding to define a 'best-site' in each of the VEs across the VPLS network, we will be drastically reducing the DF transition period as no CPU cycles will need to be spent destroying and creating new pseudowires during failover events.

5. Remote mac-flush requirement:

Having a permanent pseudowire setup would not be that effective if we end up relying solely on the current implicit mac-flush mechanism. MAC addresses are automatically aged out when the pseudowire over which they are learned is deleted. This approach would collide with the proposed 'best-site' feature, in which pseudowires are kept established on a permanent basis.

An explicit-mac-flush capable implementation would ensure that MAC-to-pseudowire bindings are cleared the moment in which a DF transition is initiated. In scenarios where 'best-site' feature is enabled, no core-facing PW will be ever torn down, so previously learned MAC entries could potentially end up pointing to an invalid PW.

Thereby, to avoid potential traffic blackholes, any successful 'best-site' implementation should be capable of supporting the explicit-mac-flush mechanism depicted in [I-D.ietf-l2vpn-vpls-multihoming draft]. F-bit was introduced in the Control-Flags bit-vector, to provide a deterministic method in which any given PE can request a remote PE to flush those mac-entries learned from the former one.

Control Flags Bit Vector

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|D|A|F|B|Z|Z|C|S| (Z = MUST Be Zero)
+-+-+-+-+-+-+-+-+
```

When making use of this feature, a DF PE will set the 'F' bit, whereas an n-DF one will clear it when sending BGP MH advertisements. A state transition from one to zero for the 'F' bit, will be interpreted by a remote PE as an indication to flush all the MACs learned from the PE that is transitioning from DF to n-DF.

6. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271].

7. IANA Considerations

8. References

8.1. Normative References

[I-D.ietf-l2vpn-vpls-multihoming] Kothari, B., Kompella, K., Henderickx, W., Balus, F., Uttaro, J., Palislamovic, S., and W. Lin, "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming-07, May 2015.

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

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service(VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

8.2. Informative References

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, May 2012.

9. Author's Addresses

Anshu Verma
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: anshuverma@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: jdrake@juniper.net

Rodny Molina
Ericsson Inc.
100 Headquarters Dr,
San Jose, CA 95134

Email: rodny.molina.maldonado@ericsson.com

Wen Lin
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: wlin@juniper.net