Network Working Group
Internet-Draft
Updates: 4761 (if approved)
Intended status: Standards Track
Expires: May 2, 2009

B. Kothari
K. Kompella
Juniper Networks
T. Spencer
AT&T

October 29, 2008

Automatic Generation of Site IDs for Virtual Private LAN Service
draft-kothari-l2vpn-auto-site-id-01.txt

Status of this Memo

This Internet-Draft will expire on May 2, 2009.

Abstract

This document defines procedures that allow for Virtual Private LAN Service (VPLS) provider edge (PE) devices that use BGP in the control plane to automatically generate VE ID values in a consistent manner.

Table of Contents

1. Introduction

Service providers are actively deploying VPLS in their networks. [RFC4761] describes mechanisms that allow VPLS PEs to use the BGP protocol to automatically discover PE membership in VPLS domains and to signal pseudowires required to carry VPLS traffic. These mechanisms make VPLS much easier to deploy and manage compared to when manual configuration of a full mesh of pseudowires is required.

A VPLS domain is an instantiation of an emulated LAN. A VPLS domain consists of a number of VPLS instances, one on each PE to which a customer site of the VPLS domain is attached. For each VPLS instance on a given PE, [RFC4761] requires a service provider to configure a route distinguisher (RD), a route target (RT) that identifies the VPLS domain, and a VE ID that is used to uniquely identify the VPLS site in the VPLS domain. The VE IDs configured must generally be unique per VPLS domain. The exception to having unique VE IDs in a VPLS domain is when a particular VPLS site is multi-homed to two or more PEs. Having the same VE ID on all the PEs to which the site is attached can be used to indicate multi-homing.

Site IDs are used by a VPLS PE to index into label blocks in order to derive the transmit and receive pseudowire labels for the pseudowires needed for transport of VPLS traffic. Thus, it is desirable for VE ID allocation in each VPLS domain to be in dense clusters in order to both minimize the number of label blocks advertised per VPLS domain and to maximize the usage of labels in a label block. For example, the set of VE IDs 1, 2, 3, 4, 5, 101, 102, 103, 104 and 105 is an efficient allocation of VE IDs for a VPLS domain.

This document describes procedures by which PEs that are offering VPLS service using BGP, as described in [RFC4761], can automatically generate VE IDs in dense clusters, thereby easing the burden on the service provider to configure and manage VE IDs per VPLS domain. The procedures to automatically generate route distinguishers for VPLS or IP VPN [RFC4364] instances on each PE are already well known. Thus, from a control plane perspective, a service provider is only required to provision route targets for each VPLS domain.

The procedures are designed to be backward compatible with the current approach of explicitly configured VE IDs. In other words, it is perfectly acceptable to have some sites in a VPLS domain with explicitly configured VE IDs (for example, to indicate multi-homing), while others have their VE IDs automatically generated. The procedures are also designed to work not only in a single autonomous system, but also in scenarios where VPLS domains span multiple autonomous systems.

2. Terminology

Terminology as described in [RFC4761] is used in this document. Additional terms are describe below.

VPLS domain:	A VPLS domain represents a bridging domain per customer. A Route Target community as described in [RFC4360] is used to identify all the PE routers participating in a particular VPLS domain.
VPLS site:	A VPLS site is a set of attachment circuits (ports) on a PE that belong to the same VPLS domain. Sites are referred to as local or remote depending on whether they belong to the PE router in context or to one of the remote PE routers (network peers).
Site ID:	VE ID as encoded in VPLS BGP NLRI.
Real advertisement:	A VPLS BGP NLRI that contains a non-zero value for VE ID, VE block offset and VE block size. Information contained in a real advertisement is required to bring up pseudowires between PEs in the same VPLS domain.
Claim advertisement:	A VPLS BGP NLRI that contains VE block offset of zero and VE block size of zero. Information contained in a claim advertisement is insufficient to bring up pseudowires between PEs in the same VPLS domain.

2.1. 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 [RFC2119].

3. Solution

The central idea of this solution is that within a VPLS domain, each PE that is configured for automatic negotiation of site identifiers will allocate an unused site ID for the VPLS site configured on the PE. It accomplishes this by keeping track of all site IDs used

within the VPLS domain based on received advertisements. A Route Target community (RT), as described in [RFC4360], is used to identify all the PE routers participating in a particular VPLS domain. When the PE needs a new site ID for an RT, it picks one of the unused site IDs for that RT and tries to claim it for a certain time period by a "claim advertisement" for this site ID. A collision occurs if this PE receives another advertisement within this time period with the same site ID and route target.

If no collision occurs, the PE takes ownership of the claimed site ID by a real advertisement for that site ID, and begins using it (establishing pseudowires, etc.). In case of a collision, the PE runs procedures as outlined in this document to resolve the collision, whereby it would either use the site ID or pick a new one based on whether it won or lost the collision resolution.

It is expected that the chance of a collision is small, especially if optimizations such as those described in this document are implemented.

The following sections provide details on the procedures required for automatic negotiation of site IDs among PEs participating in a VPLS domain.

3.1. Keeping track of site IDs in use

When a PE comes up, it SHOULD wait for time period T1 to receive advertisements from all other PEs participating in the same VPLS domain. Similarly, when a new VPLS site is configured, it SHOULD wait for time period T2 to receive advertisements from all other PEs participating in the same VPLS domain. For either of these, a PE may use BGP route refresh as described in [RFC2918] or other similar mechanisms.

The time to wait (T1 or T2) to receive relevant information can be based on configurable timers in addition to implementation specific heuristics. For example, if BGP End-of-RIB marker functionality as described in [RFC4724] is in use, the PE knows exactly when it has received all VPLS advertisements after it has come up.

A PE synthesizes a list of site IDs in use per route target from the advertisements it receives from other PEs. Note that maintaining a list of site IDs in use within a VPLS domain adds very little extra state on the VPLS PE since a VPLS PE is required to keep all VPLS advertisements for configured route targets.

When all advertisements for a site with a particular site ID are withdrawn, the site ID is deemed to be no longer in use, and MUST be

removed from the list of used site IDs. Note that a PE may advertise a single site using multiple advertisements with each advertisement carrying the label block to be used by a different set of remote sites in the VPLS domain. If the site is multi-homed, multiple PEs may advertise the same site ID.

3.2. Claiming an unused site ID

When either timer T1 or T2 expires, a PE SHOULD pick an unused site ID based on the list of site IDs in use within that VPLS domain. To indicate its interest in the selected site ID, a PE MUST send a "claim advertisement" for this site ID to all other PEs participating in the same VPLS domain. A VPLS BGP NLRI, as described in [RFC4761], for a claim advertisement MUST contain a site ID that a PE is interested in, a block offset of zero and a block size of zero.

Optimizations are possible to reduce the chance of collisions when selecting an unused site ID. For example, an implementation could determine a subset of unused site IDs and randomly select one of them instead of always selecting the unused site ID with the lowest value. An implementation might also decide to use heuristics designed to minimize the number of label blocks per VPLS domain by selecting an unused site ID that falls in the range of VPLS site advertisements from other PEs.

3.3. Collision Detection

After a claim advertisement has been send to all other PEs in a VPLS domain, a PE SHOULD wait for time period T3 to detect any collision of site ID.

A collision occurs if a PE receives any advertisement that contains the same site ID that was send in the claim advertisement. In case of a collision, a PE MUST follow the collision resolution procedures described in Section 3.4.

If a PE does not receive any advertisement with the same site ID from other PEs within time interval T3, the PE can then start using the site ID it claimed for "real advertisements". A BGP VPLS NLRI for a real advertisement MUST contain a valid site ID, a non-zero block offset and a non-zero block size in addition to valid label base value. Note that a claim advertisement contains a block offset and a block size of zero. Thus, a PE MUST withdraw a claim advertisement it previously advertised after it has send a real advertisement for its site ID.

Even after a PE starts using the site ID, it is still possible for it to receive advertisements from other PEs using the same site ID.

This can happen for example, if two PEs are trying to claim the same site ID, but neither has received each other's claim advertisements within interval interval T3. The collision resolution procedures described in Section 3.4 will handle this case as well.

3.4. Resolving a Collision

When a PE that is using a site ID or trying to claim it for its use detects a collision, it MUST run collision resolution procedures to resolve the collision. The collision could be caused by another PE using the site ID or trying to claim the site ID for its use. The collision is resolved in favor of the PE that has a better site advertisement.

Two new control flags are defined in Section 3.4.1 and collision resolution rules are described in Section 3.4.2

3.4.1. New control flags in a VPLS site advertisement

Two new control flags are proposed in this document.

1. 'A' (Automatic): Indicates whether an advertisement is for a site with explicit site ID configuration or with automatic site ID negotiation. The bit MUST be set to one in both a claim and a real advertisement for a site that is negotiating site ID by using procedures described in this document, otherwise the bit MUST be set to zero.

2. 'D' (Down): Indicates connectivity status between a CE site and a VPLS PE. The bit MUST be set to one if all the attachment circuits connecting a CE site to a VPLS PE are down.

Figure 1 shows the position of 'A' and 'D' bits in the control flags field.

Control Flags Bit Vector

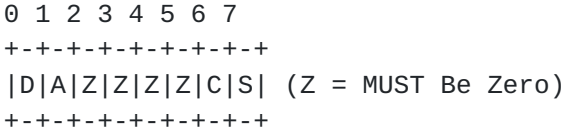

Figure 1

3.4.2. Collision Resolution Rules

The algorithm to compare two site advertisements with the same site ID is as follows:

1. An advertisement with the 'A' bit clear in the control flags is always better than an advertisement with the A bit set to one.

2. For advertisements with the same A bit value, a real advertisement is always better than a claim advertisement.

3. Between real advertisements, an advertisement with the higher BGP local preference is better. Similarly, between claim advertisements, an advertisement with the higher BGP local preference is better.

4. For advertisements with the same BGP local preference value, an advertisement with the lowest BGP next hop value is better.

The rules above MUST be applied in the order specified. The rules will pick the best advertisement in a deterministic manner, and the PE that has the best advertisement will end up using the site ID.

The PEs that have worse advertisements must withdraw all advertisements for their site ID and SHOULD try to claim another unused site ID for their use. Optimizations are possible to reduce chance of further collisions. It is recommended that a PE that lost a collision resolution wait for a short time period before making an attempt to claim another site ID. A PE that has experienced multiple collisions in succession could consider to select an unused site ID that does not fall within the range of other PEs label block advertisements, even though this would potentially result in the expansion of existing label blocks from other PEs.

3.5. Interaction with BGP path selection

In order to avoid unnecessary interaction between the rules in the previous sections and those of BGP path selection, it is recommended that PEs use unique route distinguishers (RDs) when advertising VPLS site information. This will prevent BGP route reflectors (RRs) from inadvertently filtering VPLS site advertisement information that the PEs need to receive. The procedures to automatically generate unique route distinguishers for VPLS or IP VPN [RFC4364] instances on each PE are already well known.

3.6. Lifetime of a claimed site ID

The lifetime of a claimed site ID is the duration for which the site is being advertised by the PE using a real advertisement. In other words, if all advertisements for a site that is allocated an automatically generated site ID are withdrawn, the site ID is implicitly released.

If all interfaces that connect a VPLS site to a PE are down, VPLS requires that the PE signal this event to other PEs by either withdrawing all site advertisements, or by some other means. Since withdrawing all site advertisements for a site with an automatically generated site ID has a side effect of relinquishing the site ID, it is RECOMMENDED to re-advertise the site advertisements with a "down" bit ('D' bit) set to one in the control flags instead of withdrawing the advertisements. Upon receiving a site advertisement with the 'D' bit set to one, a PE SHOULD remove all pseudowires to the advertising PE for this site ID, but MUST still consider the site ID in the advertisement to be in use. However since the new 'D' bit will not be understood by older implementations, a configurable knob should be provided in newer implementations for backward compatibility to force the withdrawal of site advertisements when all attachment circuits connecting a site to the PE go down.

3.7. Graceful Restart

When graceful restart procedures are in use for VPLS, a restarting PE SHOULD select the same site ID that it used before the restart. If a PE selects a different site ID than the one it used before the restart, VPLS forwarding will be disrupted as all other PEs within the same VPLS domain will bring down the pseudowires that were established based on the old site ID.

Without graceful restart, a restarting PE SHOULD use the procedures defined in this document for site ID negotiation.

3.8. Timers used in this approach

The procedures described in this document relies on three timers. It is RECOMMENDED that these timers be configurable. A brief description of each timer along with default values for each is given below. Note that the default values are worst case values and as such the corresponding events could be triggered by implementation specific heuristics even before the corresponding timers expire.

- T1: time to wait at startup to receive all VPLS information for configured route targets from other PEs. If End-of-RIB marker functionality [RFC4724] is in use, the PE could terminate its wait

earlier than T1 if it receives End-of-RIB markers for VPLS NLRI from all other BGP peers. The default value for T1 is recommended to be 2 minutes.

- T2: time to wait to receive VPLS information for a newly configured route target or a newly configured site for automatic site ID negotiation. The default value for T2 is recommended to be 20 seconds.

- T3: time to wait after issuing a "claim advertisement" before the PE can start using the site ID if it does not hear a competing claim. If the PE hears a competing claim within this time interval, it runs collision resolution procedures. The default value for T3 is recommended to be 30 seconds.

With the default values as specified above, a newly configured site on an operational PE will take T2 + T3 = 50 seconds before it can claim a site ID and start using it, assuming that there are no collisions with other PEs trying to use the same site ID. Similarly a PE that is restarting will take T1 + T3 = 150 seconds in the worst case before it can claim site IDs for all its sites and start using them, assuming that there are no collisions with other PEs.

3.9. Interoperating with explicitly configured site IDs

The solution presented in this document allows for automatic generation of unique site IDs per VPLS domain. However the service provider might want to explicitly configure site IDs in the network for the following reasons:

- When a site is multi-homed to two or more PEs, then the site MUST be configured with the same site ID on all the PEs. Since the solution presented here always generates unique site IDs, a service provider has to explicitly configure site IDs for multi-homed sites.

- The service provider might have PEs in the network that do not support the functionality for automatic generation of site IDs. This could be because the service provider network is multi-vendor and one or more vendors do not support this functionality, or because some PEs have older versions of software running on them that do not support the new functionality.

The procedures described in this document for sites with automatic negotiation of site IDs will interoperate with sites with explicit site configuration. Sites with explicitly configured site IDs will always use the site ID as configured. For example, if a PE detects that one of the automatically generated site IDs that it is already

using (or in the process of claiming) conflicts with an explicit site ID, it will stop using the site ID by withdrawing all advertisements with this site ID and try to claim another available site ID for its use. This behavior is achieved by the first rule in the site advertisement comparison algorithm that mandates that site advertisements with explicitly configured site IDs always win over site advertisements with automatically generated site IDs. In order for PEs to distinguish between explicitly configured and automatically generated site IDs, PEs that use automatically generated site IDs MUST set a new 'A' bit in the control flags bit-vector in advertisements for sites using automatically generated site IDs.

The compatibility of this approach with implementations that do not support this functionality also relies on these implementations ignoring claim advertisements. As explained in Section 3.2, a VPLS BGP NLRI for a claim advertisements contains a label block size of zero and a block offset of zero. An implementation that do not support automatic negotiation of site IDs SHOULD ignore an advertisement that has either a block size of zero or a block offset of zero. A claim advertisement does not have any valid labels advertised with it due to the block size of zero. In addition the block offset of zero is an invalid block offset for a real advertisement. Thus, a claim advertisement cannot be used by a remote PE to bring up a pseudowire to the site advertising the claim.

Also see the note on the 'D' bit at the end of Section 3.6.

3.10. Operation over a Multi-AS/Multi-provider MPLS core

The solution presented in this document works over a multi-AS and multi-provider core.

Section 3.4 in [RFC4761] describes three methods (a, b and c) to connect sites in a VPLS to PEs that are across multiple AS. Since VPLS advertisements in method (a) do not cross AS boundaries, procedures defined in this document for automatic site ID work in exactly the same manner as they work within an AS. In methods (b) and (c), the VPLS advertisement do cross AS boundaries, but the VE ID contained in the VPLS BGP NLRI is not changed by the intermediate ASBRs or RRs if any. Thus, each VPLS PE will receive site advertisements from all other PEs for each VPLS domain, and as described in this document, each PE will synthesize a list of site IDs in use per VPLS domain based on the remote PEs site advertisements. In such scenarios, since the advertisements have to cross AS boundaries, it is recommended to increase the time that a PE waits to hear advertisements from other PEs, both when it starts up (T1, T2) and when it waits to hear competing claims after it has

issued a claim of its own (T3).

Note that this solution only ensures that automatically generated site IDs are unique across AS boundaries. However, managing uniqueness of explicitly configured site IDs across multiple AS is outside the scope of this document.

4. Security Considerations

The procedures defined in this document allow VPLS PEs to automatically generate site IDs per VPLS domain without the service provider having to explicitly configure them. As such, no new security issues are raised beyond those that already exist in networks that use BGP-4 for exchanging VPN and VPLS membership and signaling information. Moreover, since real advertisements have priority over claim advertisements, the procedures defined in this document do not introduce new means of disrupting VPLS traffic.

5. IANA Considerations

At this time, this memo includes no request to IANA.

6. Acknowledgments

The authors would like to thank Chaitanya Kodeboyina, Nischal Sheth, Yakov Rekhter, Amit Shukla and others for the useful discussions on the subject, their review and comments.

7. References

7.1. Normative References

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

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

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

7.2. Informative References

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

Authors' Addresses

Bhupesh Kothari
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: bhupesh@juniper.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

Thomas Spencer
AT&T

Email: tsiv@att.com

Full Copyright Statement

Intellectual Property