Network Working Group
Internet Draft
Expiration Date: August 2006
Updates: RFC4364

Eric C. Rosen
Peter Psenak
Padma Pillay-Esnault
Cisco Systems, Inc.

February 2006

OSPF as the Provider/Customer Edge Protocol for BGP/MPLS IP VPNs

draft-ietf-l3vpn-ospf-2547-06.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of 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.

Abstract

Many Service Providers offer Virtual Private Network ("VPN") services to their customers, using a technique in which customer edge routers ("CE routers") are routing peers of provider edge routers ("PE routers"). The Border Gateway Protocol ("BGP") is used to distribute the customer's routes across the provider's IP backbone network, and Multiprotocol Label Switching ("MPLS") is used to tunnel customer packets across the provider's backbone. This is known as a "BGP/MPLS IP VPN". The base specification for BGP/MPLS IP VPNs presumes that the routing protocol on the interface between a PE router and a CE router is BGP. This document extends that specification by allowing the routing protocol on the PE/CE interface to be the Open Shortest

Path First ("OSPF") protocol.

This document updates draft-ietf-l3vpn-rfc2547bis-03.txt.

Table of Contents

1 Specification of Requirements 3
2 Introduction 3
3 Requirements 4
4 BGP/OSPF Interaction Procedures for PE routers 6
4.1 Overview 6
4.1.1 VRFs and OSPF Instances 6
4.1.2 VRFs and Routes 6
4.1.3 Inter-Area, Intra-Area, and External Routes 7
4.1.4 PEs and OSPF Area 0 9
4.1.5 Prevention of Loops 9
4.2 Details 10
4.2.1 Independent OSPF Instances in PEs 10
4.2.2 Router ID 10
4.2.3 OSPF Areas 10
4.2.4 OSPF Domain Identifiers 11
4.2.5 Loop Prevention 12
4.2.5.1 The DN Bit 12
4.2.5.2 Use of OSPF Route Tags 13
4.2.5.3 Other Possible Loops 14
4.2.6 Handling LSAs from the CE 14
4.2.7 Sham Links 17
4.2.7.1 Intra-Area Routes 17
4.2.7.2 Creating Sham Links 18
4.2.7.3 OSPF Protocol on Sham Links 18
4.2.7.4 Routing and Forwarding on Sham Links 19
4.2.8 VPN-IPv4 routes received via BGP 20
4.2.8.1 External Routes 21
4.2.8.2 Summary Routes 22
4.2.8.3 NSSA Routes 23
5 IANA Considerations 23
6 Security Considerations 23
7 Acknowledgments 24
8 Authors' Addresses 24
9 Normative References 25
10 Informative References 25
11 Full Copyright Statement 25
12 Intellectual Property 26

1. Specification of Requirements

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].

2. Introduction

[VPN] describes a method by which a Service Provider (SP) can use its IP backbone to provide a VPN ("Virtual Private Network") service to customers. In that method, a customer's edge devices ("CE" devices) are connected to the provider's edge routers ("PE" routers). If the CE device is a router, then the PE router may become a routing peer of the CE router (in some routing protocol), and may as a result learn the routes which lead to the CE's site and which need to be distributed to other PE routers that attach to the same VPN.

The PE routers which attach to a common VPN use BGP ("Border Gateway Protocol") to distribute the VPN's routes to each other. A CE router can then learn the routes to other sites in the VPN by peering with its attached PE router in a routing protocol. CE routers at different sites do not, however, peer with each other.

It can be expected that many VPNs will use OSPF ("Open Shortest Path First") as their IGP ("Interior Gateway Protocol"), i.e., the routing protocol used by a network for the distribution of internal routes within that network). This does not necessarily mean that the PE routers need to use OSPF to peer with the CE routers. Each site in a VPN can use OSPF as its intra-site routing protocol, while using, e.g., BGP or RIP ("Routing Information Protocol") to distribute routes to a PE router. However, it is certainly convenient, when OSPF is being used intra-site, to use it on the PE-CE link as well, and [VPN] explicitly allows this.

Like anything else, the use of OSPF on the PE-CE link has advantages and disadvantages. The disadvantage to using OSPF on the PE-CE link is that it gets the SP's PE router involved, however peripherally, in a VPN site's IGP. The advantages though are:

- The administrators of the CE router need not have any expertise in any routing protocol other than OSPF.

- The CE routers do not need to have support for any routing protocols other than OSPF.

- If a customer is transitioning his network from a traditional OSPF backbone to the VPN service described in [VPN], the use of OSPF on the PE-CE link eases the transitional issues.

It seems likely that some SPs and their customers will resolve these trade-offs in favor of the use of OSPF on the PE-CE link. Thus we need to specify the procedures which must be implemented by a PE router in order to make this possible. (No special procedures are needed in the CE router though; CE routers just run whatever OSPF implementations they may have.)

3. Requirements

Consider a set of VPN sites which are thought of as being in the same "OSPF domain". Two sites are considered to be in the same OSPF domain if it is intended that routes from one site to the other be considered to be intra-network routes. A set of OSPF sites in the same domain will almost certainly be a set of sites which together constitute an "intranet", and each of which runs OSPF as its intra-site routing protocol.

Per [VPN], the VPN routes are distributed among the PE routers by BGP. If the PE uses OSPF to distribute routes to the CE router, the standard procedures governing BGP/OSPF interactions [OSPFv2] would cause routes from one site to be delivered to another in type 5 LSAs ("Link State Advertisements"), as "AS-external" routes. This is undesirable; it would be much better to deliver such routes in type 3 LSAs (as inter-area routes), so that they can be distinguished from any "real" AS-external routes that may be circulating in the VPN. (That is, so that they can be distinguished by OSPF from routes which really do not come from within the VPN.) Hence it is necessary for the PE routers to implement a modified version of the BGP/OSPF interaction procedures.

In fact, we would like to have a very general set of procedures which allows a customer to easily replace a legacy private OSPF backbone with the VPN service. We would like this procedure to meet the following set of requirements:

- The procedures should not make assumptions about the OSPF topology. In particular, it should not be assumed that customer sites are OSPF stub sites or NSSA ("Not So Stubby Area") sites. Nor should it be assumed that a customer site contains only one OSPF area, or that it has no area 0 routers.

- If VPN sites A and B are in the same OSPF domain, then routes from one should be presented to the other as OSPF intra-network routes. In general, this can be done by presenting such routes as inter-area routes, in type 3 LSAs.

  Note that this allows two VPN sites to be connected via an "OSPF backdoor link". That is, one can have an OSPF link between the two sites which is used only when the VPN backbone is unavailable. (This would not be possible with the ordinary BGP/OSPF interaction procedures. The ordinary procedures would present routes via the VPN backbone as AS-external routes, and these could never be preferred to intra-network routes.) This may be very useful during a period of transition from a legacy OSPF backbone to a VPN backbone.

- It should be possible to make use of an "OSPF backdoor link" between two sites, even if the two sites are in the same OSPF area, and neither of the routers attached to the inter-site backdoor link is an area 0 router. This can also be very useful during a transition period, and eliminates any need to reconfigure the sites' routers to be ABRs ("Area Border Routers").

  Assuming that it is desired to have the route via the VPN backbone be preferred to the backdoor route, the VPN backbone itself must be presented to the CE routers at each site as a link between the two PE routers to which the CE routers are respectively attached.

- CE routers, connected to PE routers of the VPN service, may themselves function as OSPF backbone (area 0) routers. An OSPF backbone may even consist of several "segments" which are interconnected themselves only via the VPN service. In such a scenario, full intercommunication between sites connected to different segments of the OSPF backbone should still be possible.

- The transition from the legacy private OSPF backbone to the VPN service must be simple and straightforward. The transition is likely to be phased, such that customer sites are migrated one by one from the legacy private OSPF backbone to the VPN service. During the transition, any given site might be connected to the VPN service, to the legacy OSPF backbone, or to both. Complete connectivity among all such sites must be maintained.

  Since the VPN service is to replace the legacy backbone, it must be possible, by suitable adjustment of the OSPF metrics, to make OSPF prefer routes which traverse the SP's VPN backbone to alternative routes which do not.

- The OSPF metric assigned to a given route should be carried transparently over the VPN backbone.

Routes from sites which are not in the same OSPF domain will appear as AS-external routes.

We presuppose familiarity with the contents of [OSPFv2], including the OSPF LSA types, and will refer without further exegesis to type 1, 2, 3, etc., LSAs. Familiarity with [VPN] is also presupposed.

4. BGP/OSPF Interaction Procedures for PE routers

4.1. Overview

4.1.1. VRFs and OSPF Instances

A PE router which attaches to more than one OSPF domain MUST run an independent instance of OSPF for each domain. If the PE is running OSPF as its IGP ("Interior Gateway Protocol"), the instance of OSPF running as the IGP must be separate and independent from any other instance of OSPF which the PE is running. (Whether these instances are realized as separate processes, or merely as separate contexts of a common process, is an implementation matter.) Each interface that attaches to a VPN site belongs to no more than one OSPF instance.

[VPN] defines the notion of a Per-Site Routing and Forwarding Table, or VRF. Each VRF is associated with a set of interfaces. If a VRF is associated with a particular interface, and that interface belongs to a particular OSPF instance, then that OSPF instance is said to be associated with the VRF. If two interfaces belong to the same OSPF instance, then both interfaces must be associated with the same VRF.

If an interface attaches a PE to a CE, and that interface is associated with a VRF, we will speak of the CE as being associated with the VRF.

4.1.2. VRFs and Routes

OSPF is used to distribute routes from a CE to a PE. The standard OSPF decision process is used to install the best OSPF-distributed routes in the VRF.

Per [VPN], BGP is used to distribute VPN-IPv4 routes among PE routers. An OSPF route installed in a VRF may be "exported" by being redistributed into BGP as a VPN-IPv4 route. It may then be distributed by BGP to other PEs. At the other PEs, a VPN-IPv4 route

may "imported" by a VRF, and may then be redistributed into one or more of the OSPF instances associated with that VRF.

Import from and export to particular VRFs is controlled by the use of the Route Target Extended Communities attribute (or more simply, "Route Target" or "RT"), as specified in [VPN].

A VPN-IPv4 route is "eligible for import" into a particular VRF if its Route Target is identical to one of the VRF's import Route Targets. The standard BGP decision process is used to select, from among the routes eligible for import, the set of VPN-IPv4 routes to be "installed" in the VRF.

If a VRF contains both an OSPF-distributed route and a VPN-IPv4 route for the same IPv4 prefix, then the OSPF-distributed route is preferred. In general, this means that forwarding is done according to the OSPF route. The one exception to this rule has to do with the "sham link". If the next hop interface for an installed (OSPF-distributed) route is the sham link, forwarding is done according to a corresponding BGP route. This is detailed in section 4.2.7.4.

To meet the requirements of section 3, a PE that installs a particular route into a particular VRF needs to know whether that route was originally an OSPF route, and if so, whether the OSPF instance from which it was redistributed into BGP is in the same domain as the OSPF instances into which the route may be redistributed. Therefore a domain identifier is encoded as a BGP Extended Communities attribute [EXTCOMM], and distributed by BGP along with the VPN-IPv4 route. The route's OSPF metric and OSPF route type are also carried as BGP attributes of the route.

4.1.3. Inter-Area, Intra-Area, and External Routes

If a PE installs a particular VPN-IPv4 route (learned via BGP) in a VRF, and if this is the preferred BGP route for the corresponding IPv4 prefix, then the corresponding IPv4 route is then "eligible for redistribution" into each OSPF instance that is associated with the VRF. As a result, it may be advertised to each CE in an LSA.

Whether a route which is eligible for redistribution into OSPF is actually redistributed into a particular OSPF instance may depend upon the configuration. For instance, the PE may be configured to only distribute the default route into a given OSPF instance. In this case, the routes which are eligible for redistribution would not actually be redistributed.

In the following, we discuss the procedures for redistributing a BGP-distributed VPN-IPv4 route into OSPF; these are the procedures to be followed whenever such a route is eligible to be redistributed into OSPF and the configuration does not prevent such redistribution.

If the route is from a different OSPF domain than OSPF instance into which it is being redistributed, or if the route is not from an OSPF domain at all, then the route is considered to be an external route.

If the route is from the same OSPF domain as the OSPF instance into which it is being redistributed, and if it was originally advertised to a PE as an OSPF external route or an OSPF NSSA route, it will be treated as an external route. Following the normal OSPF procedures, external routes may be advertised to the CE in type 5 LSAs, or in type 7 LSAs, or not at all, depending on the type of area to which the PE/CE link belongs.

If the route is from the same OSPF domain as the OSPF instance into which it is being redistributed, and and if it was originally advertised to a PE as an inter-area or intra-area route, the route will generally be advertised to the CE as an inter-area route (in a type 3 LSA).

As a special case, suppose PE1 attaches to CE1, and PE2 attaches to CE2, where:

- the OSPF instance containing the PE1-CE1 link and the OSPF instance containing the PE2-CE2 link are in the same OSPF domain, and

- the PE1-CE1 and PE2-CE2 links are in the same OSPF area A (as determined by the configured OSPF area number),

then PE1 may flood to CE1 a type 1 LSA advertising a link to PE2, and PE2 may flood to CE2 a type 1 LSA advertising a link to PE1. The link advertised in these LSAs is known as a "sham link", and it is advertised as a link in area A. This makes it look to routers within area A as if the path from CE1 to PE1 across the service provider's network to PE2 to CE2 is an intra-area path. Sham links are an OPTIONAL feature of this specification, and are used only when it is necessary to have the service provider's network treated as an intra-area link. See section 4.2.7 for further details about the sham link.

The precise details by which a PE determines the type of LSA used to advertise a particular route to a CE are specified in section 4.2.8. Note that if the VRF is associated with multiple OSPF instances, the type of LSA used to advertise the route might be different in

different instances.

Note that if a VRF is associated with several OSPF instances, a given route may be redistributed into some or all of those OSPF instances, depending on the characteristics of each instance. If redistributed into two or more OSPF instances, it may be advertised within each instance using a different type of LSA, again depending on the characteristics of each instance.

4.1.4. PEs and OSPF Area 0

Within a given OSPF domain, a PE may attach to multiple CEs. Each PE/CE link is assigned (by configuration) to an OSPF area. Any link can be assigned to any area, including area 0.

If a PE attaches to a CE via a link which is in a non-zero area, then the PE serves as an ABR for that area.

PEs can thus be considered to be OSPF "area 0 routers", i.e., they can be considered to be part of the "OSPF backbone". Thus they are allowed to distribute inter-area routes to the CE via Type 3 LSAs.

If the OSPF domain has any area 0 routers other than the PE routers, then at least one of those MUST be a CE router, and MUST have an area 0 link to at least one PE router. This adjacency MAY be via an OSPF virtual link. (The ability to use an OSPF virtual link in this way is an OPTIONAL feature.) This is necessary to ensure that inter-area routes and AS-external routes can be leaked between the PE routers and the non-PE OSPF backbone.

Two sites which are not in the same OSPF area will see the VPN backbone as being an integral part of the OSPF backbone. However, if there are area 0 routers which are NOT PE routers, then the VPN backbone actually functions as a sort of higher level backbone, providing a third level of hierarchy above area 0. This allows a legacy OSPF backbone to become disconnected during a transition period, as long as the various segments all attach to the VPN backbone.

4.1.5. Prevention of Loops

If a route sent form a PE router to a CE router could then be received by another PE router from one of its own CE routers, it would be possible for routing loops to occur. To prevent this, a PE sets the DN bit [OSPF-DN] in any LSA that it sends to a CE, and a PE ignores any LSA received from a CE which already has the DN bit sent.

Older implementations may use an OSPF Route Tag instead of the DN bit, in some cases. See sections 4.2.5.1 and 4.2.5.2.

4.2. Details

4.2.1. Independent OSPF Instances in PEs

The PE MUST support one OSPF instance for each OSPF domain to which it attaches. These OSPF instances function independently and do not leak routes to each other. Each instance of OSPF MUST be associated with a single VRF. If n CEs associated with that VRF are running OSPF on their respective PE/CE links, then those n CEs are OSPF adjacencies of the PE in the corresponding instance of OSPF.

Generally, though not necessarily, if the PE attaches to several CEs in the same OSPF domain, it will associate the interfaces to those PEs with a single VRF.

4.2.2. Router ID

If a PE and a CE are communicating via OSPF, the PE will have an OSPF Router ID which is valid (i.e., unique) within the OSPF domain. More precisely, each OSPF instance has a Router ID. Different OSPF instances may have different Router IDs.

4.2.3. OSPF Areas

A PE-CE link may be in any area, including area 0; this is a matter of the OSPF configuration.

If a PE has a link which belongs to a non-zero area, the PE functions as an Area Border Router (ABR) for that area.

PEs do not pass along the link state topology from one site to another (except in the case where a sham link is used, see section 4.2.7).

Per [OSPF, section 3.1], "the OSPF backbone always contains all area border routers". The PE routers are therefore considered to be area 0 routers. Section 3.1 of [OSPFv2] also requires that area 0 be contiguous. It follows that if the OSPF domain has any area 0 routers other than the PE routers, at least one of those MUST be a CE router, and it MUST have an area 0 link (possibly a virtual link) to at least one PE router.

4.2.4. OSPF Domain Identifiers

Each OSPF instance MUST be associated with one or more Domain Identifiers. This MUST be configurable, and the default value (if none is configured) SHOULD be NULL.

If an OSPF instance has multiple Domain Identifiers, one of these is considered to be its "primary" Domain Identifier; this MUST be determinable by configuration. If an OSPF instance has exactly one Domain Identifier, this is of course its primary Domain Identifier. If an OSPF instance has more than one Domain Identifiers, the NULL Domain Identifier MUST NOT be one of them.

If a route is installed in a VRF by a particular OSPF instance, the primary Domain Identifier of that OSPF instance is considered to be the route's Domain Identifier.

Consider a route R which is installed in a VRF by OSPF instance I1, then redistributed into BGP as a VPN-IPv4 route, then installed by BGP in another VRF. If R needs to be redistributed into OSPF instance I2, associated with the latter VRF, the way in which R is advertised in I2 will depend upon whether R's Domain Identifier is one of I2's Domain Identifiers. If R's Domain Identifier is not one of I2's Domain Identifiers, then if R is redistributed into I2, R will be advertised as an AS-external route, no matter what its OSPF route type is. If, on the other hand, R's Domain Identifier is one of I2's Domain Identifiers, how R is advertised will depend upon R's OSPF route type.

If two OSPF instances are in the same OSPF domain, then either:

1. They both have the NULL Domain Identifier, OR

2. Each OSPF instance has the primary Domain Identifier of the other as one of its own Domain Identifiers.

If two OSPF instances are in different OSPF domains, then either:

3. They both have the NULL Domain Identifier, OR

4. Neither OSPF instance has the Primary Domain Identifier of the other as one of its own Domain Identifiers.

(Note that if two OSPF instances each have the NULL Domain Identifier, we cannot tell from the Domain Identifier whether they are in the same OSPF Domain or not. If they are in different domains, and if routes from one are distributed into the other, the routes will appear as intra-network routes, which may not be what is

intended.)

A Domain Identifier is an eight-byte quantity which is a valid BGP Extended Communities attribute, as specified in section 4.2.4. If a particular OSPF instance has a non-NULL Domain Identifier, when routes from that OSPF instance are distributed by BGP as VPN-IPv4 routes, the routes MUST carry the Domain Identifier Extended Communities attribute which corresponds to the OSPF instance's Primary Domain Identifier. If the OSPF instance's Domain Identifier is NULL, the Domain Identifier Extended Communities attribute MAY be omitted when routes from that OSPF instance are distributed by BGP; alternatively, a value of the Domain Identifier Extended Communities attribute which represents NULL (see section 4.2.4) MAY be carried with the route.

If the OSPF instances of an OSPF domain are given one or more non-NULL Domain Identifiers, this procedure allows us to determine whether a particular OSPF-originated VPN-IPv4 route belongs to the same domain as a given OSPF instance. We can then determine whether the route should be redistributed to that OSPF instance as an inter-area route or as an OSPF AS-external route. Details can be found in sections 4.2.4 and 4.2.8.1.

4.2.5. Loop Prevention

4.2.5.1. The DN Bit

When a type 3 LSA is sent from a PE router to a CE router, the DN bit [OSPF-DN] in the LSA Options field MUST be set. This is used to ensure that if any CE router sends this type 3 LSA to a PE router, the PE router will not further redistribute it.

When a PE router needs to distribute to a CE router a route that comes from a site outside the latter's OSPF domain, the PE router presents itself as an ASBR ("Autonomous System Border Router"), and distributes the route in a type 5 LSA. The DN bit [OSPF-DN] MUST be set in these LSAs, to ensure that they will be ignored by any other PE routers that receive them.

There are deployed implementations which do not set the DN bit, but instead use OSPF route tagging to ensure that a type 5 LSA generated by a PE router will be ignored by any other PE router that may receive it. A special OSPF route tag, which we will call the VPN Route Tag (see section 4.2.5.2), is used for this purpose. To ensure backwards compatibility, all implementations adhering to this specification MUST by default support the VPN Route Tag procedures specified in sections 4.2.5.2, 4.2.8.1 and 4.2.8.2. When it is no

longer necessary to use the VPN Route Tag in a particular deployment, its use (both sending and receiving) may be disabled by configuration.

4.2.5.2. Use of OSPF Route Tags

If a particular VRF in a PE is associated with an instance of OSPF, then by default it MUST be configured with a special OSPF route tag value, which we call the VPN Route Tag. By default, this route tag MUST be included in the Type 5 LSAs which the PE originates (as the result of receiving a BGP-distributed VPN-IPv4 route, see section 4.2.8) and sends to any of the attached CEs.

The configuration and inclusion of the VPN Route Tag is required for backwards compatibility with deployed implementations that do not set the DN bit in type 5 LSAs. The inclusion of the VPN Route Tag may be disabled by configuration, if it has been determined that it is no longer needed for backwards compatibility.

The value of the VPN Route Tag is arbitrary, but must be distinct from any OSPF Route Tag being used within the OSPF domain. Its value MUST therefore be configurable. If the Autonomous System number of the VPN backbone is two bytes long, the default value SHOULD be an automatically computed tag based on that Autonomous System number:

```
Tag = <Automatic = 1, Complete = 1, PathLength = 01>

  0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |1|1|0|1|     ArbitraryTag      |       AutonomousSystem        |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

  1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 _AS number of the VPN Backbone_
```

If the Autonomous System number is four bytes long, then a Route Tag value MUST be configured, and it MUST be distinct from any Route Tag used within the VPN itself.

If a PE router needs to use OSPF to distribute to a CE router a route which comes from a site outside the CE router's OSPF domain, the PE router SHOULD present itself to the CE router as an Autonomous System Border Router (ASBR), and SHOULD report such routes as AS-external routes. That is, these PE routers originate Type 5 LSAs reporting the extra-domain routes as AS-external routes. Each such Type 5 LSA

MUST contain an OSPF route tag whose value is that of the VPN Route Tag. This tag identifies the route as having come from a PE router. The VPN Route Tag MUST be used to ensure that a Type 5 LSA originated by a PE router is not redistributed through the OSPF area to another PE router.

4.2.5.3. Other Possible Loops

The procedures specified in this document ensure that if routing information derived from a BGP-distributed VPN-IPv4 route is distributed into OSPF, it cannot be redistributed back into BGP as a VPN-IPv4 route, as long as the DN bit and/or VPN route tag is maintained within the OSPF domain. This does not eliminate all possible sources of loops. For example, if a BGP VPN-IPv4 route is distributed into OSPF, and then distributed into RIP (where all the information needed to prevent looping is lost), and then distributed back into OSPF, then it is possible that it could be distributed back into BGP as a VPN-IPv4 route, thereby causing a loop.

Therefore extreme care must be taken if there is any mutual redistribution of routes between the OSPF domain and any third (i.e., not the VPN backbone) routing domain. (If the third routing domain is a BGP domain, e.g., the public Internet, the ordinary BGP route prevention measures will prevent the route from reentering the OSPF domain.)

4.2.6. Handling LSAs from the CE

This section specifies the way in which a PE router handles the OSPF LSAs it receives from a CE router.

When a PE router receives, from a CE router, any LSA with the DN bit [OSPF-DN] set, the information from that LSA MUST NOT be used by the route calculation. If a Type 5 LSA is received from the CE, and if it has an OSPF route tag value equal to the VPN Route Tag (see section 4.2.5.2), then the information from that LSA MUST NOT be used by the route calculation.

Otherwise, the PE must examine the corresponding VRF. For every address prefix which was installed in the VRF by one of its associated OSPF instances, the PE must create a VPN-IPv4 route in BGP. Each such route will have some of the following Extended Communities attributes:

- The OSPF Domain Identifier Extended Communities attribute. If the OSPF instance which installed the route has a non-NULL primary Domain Identifier, this MUST be present; if that OSPF instance has only a NULL Domain Identifier, it MAY be omitted. This attribute is encoded with a two-byte type field, and its type is either 0005, 0105, or 0205. For backwards compatibility, the type 8005 MAY be used as well, and is treated as if it were 0005. If the OSPF instance has a NULL Domain Identifier, and the OSPF Domain Identifier Extended Communities attribute is present, then the attribute's value field must be all zeroes, and its type field may be any of 0005, 0105, 0205, or 8005.

- OSPF Route Type Extended Communities Attribute. This attribute MUST be present. It is encoded with a two-byte type field, and its type is 0306. To ensure backwards compatibility, the type 8000 SHOULD be accepted as well, and treated as if it were type 0306. The remaining six bytes of the Attribute are encoded as follows:

```
+-------+-------+-------+-------+-------+-------+
|        Area Number            | Route |Options|
|                               | Type  |       |
+-------+-------+-------+-------+-------+-------+
```

  * Area Number: 4 bytes, encoding a 32-bit area number. For AS-external routes, the value is 0. A non-zero value identifies the route as being internal to the OSPF domain, and as being within the identified area. Area numbers are relative to a particular OSPF domain.

  * OSPF Route Type: 1 byte, encoded as follows:

    ** 1 or 2 for intra-area routes (depending on whether the route came from a type 1 or a type 2 LSA).

    ** 3 for inter-area routes

    ** 5 for external routes (area number must be 0)

    ** 7 for NSSA routes.

  Note that the procedures of section 4.2.8 do not make any distinction between routes types 1, 2, and 3. If BGP installs a route of one of these types in the VRF, and if that route is selected for redistribution into OSPF, it will be advertised by OSPF in either a type 3 or a type 5 LSA,

depending on the domain identifier.

* Options: 1 byte. Currently this is only used if the route type is 5 or 7. Setting the least significant bit in the field indicates that the route carries a type 2 metric.

- OSPF Router ID Extended Communities Attribute. This OPTIONAL attribute specifies the OSPF Router ID of the system that is identified in the BGP Next Hop attribute. More precisely, it specifies the OSPF Router Id of the PE in the OSPF instance which installed the route into the VRF from which this route was exported. This attribute is encoded with a two-byte type field, and its type is 0107, with the Router ID itself carried in the first 4 bytes of the value field. The type 8001 SHOULD be accepted as well, to ensure backwards compatibility, and should be treated as if it were 0107.

- MED ("Multi_EXIT_DISC attribute"). By default, this SHOULD be set to the value of the OSPF distance associated with the route, plus 1.

The intention of all this is the following. OSPF Routes from one site are converted to BGP, distributed across the VPN backbone, and possibly converted back to OSPF routes before being distributed into another site. With these attributes, BGP carries enough information about the route to enable the route to be converted back into OSPF "transparently", just as if BGP had not been involved.

Routes which a PE receives in type 4 LSAs MUST NOT be redistributed to BGP.

The attributes specified above are in addition to any other attributes which routes must carry in accord with the [VPN].

The Site of Origin attribute, which is usually required by [VPN], is OPTIONAL for routes which a PE learns from a CE via OSPF.

Use of the Site of Origin attribute would, in the case of a multiply homed site (i.e., a site attached to several PE routers), prevent an intra-site route from being reinjected into a site from the VPN backbone. Such a reinjection would not harm the routing, because the route via the VPN backbone would be advertised in a type 3 LSA, and hence would appear to be an inter-area route; the real intra-area route would be preferred. But unnecessary overhead would be introduced. On the other hand, if the Site of Origin attribute is not used, a partitioned site will find itself automatically repaired, since traffic from one partition to the other will automatically travel via the VPN backbone. Therefore the use of a Site of Origin

attribute is optional, so that a trade-off can be made between the cost of the increased overhead and the value of automatic partition repair.

4.2.7. Sham Links

This section describes the protocol and procedures necessary for the support of "Sham Links," as defined herein. Support for sham links is an OPTIONAL feature of this specification.

4.2.7.1. Intra-Area Routes

Suppose there are two sites in the same OSPF area. Each site is attached to a different PE router, and there is also an intra-area OSPF link connecting the two sites.

It is possible to treat these two sites as a single VPN site which just happens to be multihomed to the backbone. This is in fact the simplest thing to do, and is perfectly adequate, providing that the preferred route between the two sites is via the intra-area OSPF link (a "backdoor link"), rather than via the VPN backbone. There will be routes between sites that go through the PE routers, but these routes will appear to be inter-area routes, and OSPF will consider them to be less preferable than the intra-area routes through the backdoor link.

If it is desired to have OSPF prefer the routes through the backbone over the routes through the backdoor link, then the routes through the backbone must be appear to be intra-area routes. To make a route through the backbone appear to be an intra-area route, it is necessary to make it appear as if there is an intra-area link connecting the two PE routers. This is what we refer to as a "sham link". (If the two sites attach to the same PE router, this of course is not necessary.)

A sham link can be thought of as a relation between two VRFs. If two VRFs are to be connected by a sham link, each VRF must be associated with a "Sham Link Endpoint Address", a 32-bit IPv4 address which is treated as an address of the PE router containing that VRF. The Sham Link Endpoint Address is an address in the VPN's address space, not the SP's address space. The Sham Link Endpoint Address associated with a VRF MUST be configurable. If the VRF is associated with only a single OSPF instance, and if the PE's router id in that OSPF instance is an IP address, then the Sham Link Endpoint Address MAY default to that Router ID. If a VRF is associated with several OSPF instances, each sham link belongs to a single OSPF instance.

For a given OSPF instance, a VRF needs only a single Sham Link Endpoint Address, no matter how many sham links it has. The Sham Link Endpoint Address MUST be distributed by BGP as a VPN-IPv4 address whose IPv4 address prefix part is 32 bits long. The Sham Link Endpoint Address MUST NOT be advertised by OSPF; if there is no BGP route to the Sham Link Endpoint Address, that address is to appear to be unreachable, so that the sham link appears to be down.

4.2.7.2. Creating Sham Links

Sham links are manually configured.

For a sham link to exist between two VRFs, each VRF has to be configured to create a sham link to the other, where the "other" is identified by its sham link endpoint address. No more than one sham link with the same pair of sham link endpoint addresses will ever be created. This specification does not include procedures for single-ended manual configuration of the sham link.

Note that sham links may be created for any area, including area 0.

A sham link connecting two VRFs is considered to be up if and only if a route to the 32-bit remote endpoint address of the sham link has been installed in VRF.

The sham link endpoint address MUST NOT be used as the endpoint address of an OSPF Virtual Link.

4.2.7.3. OSPF Protocol on Sham Links

An OSPF protocol packet sent on a Sham Link from one PE to another must have as its IP source address the Sham Link Endpoint Address of the sender, and as its IP destination address the Sham Link Endpoint Address of the receiver. The packet will travel from one PE router to the other over the VPN backbone, which means that it can be expected to traverse multiple hops. As such, it's TTL ("Time to Live") field must be set appropriately.

An OSPF protocol packet is regarded as having been received on a particular sham link if and only if the following three conditions hold:

- The packet arrives as an MPLS packet, and its MPLS label stack causes it to be "delivered" to the local sham link endpoint address;

- The packet's IP destination address is the local sham link endpoint address;

- The packet's IP source address is the remote sham link endpoint address.

Sham links SHOULD be treated by OSPF as OSPF Demand Circuits. This means that LSAs will be flooded over them, but periodic refresh traffic is avoided. Note that, as long as the backdoor link is up, flooding the LSAs over the sham link serves no purpose. However, if the backdoor link goes down, OSPF does not have mechanisms enabling the routers in one site to rapidly flush the LSAs from the other site. Therefore it is still necessary to maintain synchronization among the LSA databases at the two sites, hence the flooding over the sham link.

The sham link is an unnumbered point-to-point intra-area link, and is advertised as a type 1 link in a type 1 LSA.

The OSPF metric associated with a sham link MUST be configurable (and there MUST be a configurable default). Whether traffic between the sites flows via a backdoor link or via the VPN backbone (i.e., via the sham link) depends on the settings of the OSPF link metrics. The metrics can be set so that the backdoor link is not used unless connectivity via the VPN backbone fails, for example.

The default Hello Interval for sham links is 10 seconds, and the default Router Dead Interval for sham links is 40 seconds.

4.2.7.4. Routing and Forwarding on Sham Links

If a PE determines that the next hop interface for a particular route is a sham link, then the PE SHOULD NOT redistribute that route into BGP as a VPN-IPv4 route.

Any other route advertised in an LSA which is transmitted over a sham link MUST also be redistributed (by the PE flooding the LSA over the sham link) into BGP. This means that if the preferred (OSPF) route for a given address prefix has the sham link as its next hop interface, then there will also be a "corresponding BGP route", for that same address prefix, installed in the VRF. Per section 4.1.2, the OSPF route is preferred. However, when forwarding a packet, if the preferred route for that packet has the sham link as its next hop interface, then the packet MUST be forwarded according to the corresponding BGP route. That is, it will be forwarded as if the corresponding BGP route had been the preferred route. The "corresponding BGP route" is always a VPN-IPv4 route; the procedure

for forwarding a packet over a VPN-IPv4 route is described in [VPN].

This same rule applies to any packet whose IP destination address is the remote endpoint address of a sham link. Such packets MUST be forwarded according to the corresponding BGP route.

4.2.8. VPN-IPv4 routes received via BGP

This section describes how the PE router handles VPN-IPv4 routes received via BGP.

If a received BGP VPN-IPv4 route is not installed in the VRF, nothing is reported to the CE. A received route will not be installed into the VRF if the BGP decision process regards some other some other route is preferable. When installed in the VRF, the route appears to be an IPv4 route.

A BGP route installed in the VRF is not necessarily used for forwarding. If an OSPF route for the same IPv4 address prefix has been installed in the VRF, the OSPF route will be used for forwarding, except in the case where the OSPF route's next hop interface is a sham link.

If a BGP route installed in the VRF is used for forwarding, then the BGP route is redistributed into OSPF, and possibly reported to the CEs in an OSPF LSA. The sort of LSA, if any, to be generated depends on various characteristics of the BGP route, as detailed in subsequent sections of this document.

The procedure for forwarding a packet over a VPN-IPv4 route is described in [VPN].

In the following, we specify what is reported, in OSPF LSAs, by the PE to the CE, assuming that the PE is not configured to do any further summarization or filtering of the routing information before reporting it to the CE.

When sending an LSA to the CE, it may be necessary to set the DN bit. See section 4.2.5.1 for the rules regarding the DN bit.

When sending an LSA to the CE, it may be necessary to set the OSPF Route Tag. See section 4.2.5.2 for the rules about setting the OSPF Route Tag.

When type 5 LSAs are sent, the Forwarding Address is set to 0.

4.2.8.1. External Routes

With respect to a particular OSPF instance associated with a VRF, a VPN-IPv4 route which is installed in the VRF and then selected as the preferred route is treated as an External Route if one of the following conditions holds:

- the route type field of the OSPF Route Type Extended Community has an OSPF route type of "external",

- the route is from a different domain than the domain of the OSPF instance.

The rules for determining whether a route is from a different domain than a particular OSPF instance are the following. The OSPF Domain Identifier Extended Communities attribute carried by the route is compared with the OSPF Domain Identifier Extended Communities attribute(s) with which the OSPF instance has been configured (if any). In general, when two such attributes are compared, all eight bytes must be compared. Thus two OSPF Domain Identifier Extended Communities attributes are regarded as equal if and only if one of the following three conditions holds:

1. They are identical in all eight bytes.

2. They are identical in their lower order six bytes (value field), but one attribute has two high order bytes (type field) of 0005 and the other has two high order bytes (type field) of 8005. (This condition is for backwards compatibility.)

3. The lower order six bytes (value field) of both attributes consist entirely of zeroes. In this case, the two attributes are considered to be identical irrespective of their type fields, and they are regarded as representing the NULL Domain Identifier.

If a VPN-IPv4 route has an OSPF Domain Identifier Extended Communities attribute, we say that that route is in the identified domain. If the value field of the Extended Communities attribute consists of all zeroes, then the identified domain is the NULL domain, and the route is said to belong to the NULL domain. If the route does not have an OSPF Domain Identified Extended Communities attribute, then the route belongs to the NULL domain.

Every OSPF instance is associated with one or more Domain Identifiers, though possibly only with the NULL domain identifier. If an OSPF instance is associated with a particular Domain Identifier, we will say that it belongs to the identified domain.

If a VPN-IPv4 route is to be redistributed to a particular instance, it must be determined whether that route belongs and that OSPF instance belong to the same domain. A route and an OSPF instance belong to the same domain if and only if one of the following conditions holds:

1. The route and the OSPF instance each belong to the NULL domain.

2. The domain to which the route belongs is the domain to which the OSPF instance belongs. (I.e., the route's Domain Identifier is equal to the OSPF instance's domain identifier, as determined by the definitions given earlier in this section.)

If the route and the VRF do not belong to the same domain, the route is treated as an external route.

If an external route is redistributed into an OSPF instance, the route may or may not be advertised to a particular CE, depending on the configuration and on the type of area to which the PE/CE link belongs. If the route is advertised, and the PE/CE link belongs to a NSSA area, it is advertised in a type 7 LSA. Otherwise, if the route is advertised it is advertised in a type 5 LSA. The LSA will be originated by the PE.

The DN bit (section 4.2.5.1) MUST be set in the LSA. The VPN Route Tag (see section 4.2.5.2) MUST be placed in the LSA, unless the use of the VPN Route Tag has been turned off by configuration.

By default, a type 2 metric value is included in the LSA, unless the options field of the OSPF Route Type Extended Communities attribute of the VPN-IPv4 route specifies that the metric should be type 1.

By default, the value of the metric its value is taken from the MED attribute of the VPN-IPv4 route. If the MED is not present, a default metric value is used. (The default type 1 metric and the default type 2 metric MAY be different.)

Note that this way of handling external routes makes every PE appear to be an ASBR attached to all the external routes. In a multihomed site, this can result in a number of type 5 LSAs containing the same information.

4.2.8.2. Summary Routes

If a route and the VRF into which it is imported belong to the same domain, then the route should be treated as if it had been received

in an OSPF type 3 LSA. This means that the PE will report the route in a type 3 LSA to the CE. (Note that this case is possible even if the VPN-IPv4 route carries an area number identical to that of the CE router. This means that if an area is "partitioned" such that the two pieces are connected only via the VPN backbone, it appears to be two areas, with inter-area routes between them.)

4.2.8.3. NSSA Routes

NSSA routes are treated the same as external routes, as described in section 4.2.8.1.

5. IANA Considerations

Section 11 of [EXTCOMM] calls upon IANA to create a registry for BGP Extended Communities Type Field and Extended Type Field values. Section 4.2.6 of this document assigns new values for the BGP Extended Communities Extended Type Field. These values all fall within the range of values which [EXTCOMM] states "are to be assigned by IANA, using the 'First Come, First Served' policy defined in RFC 2434".

The BGP Extended Communities Extended Type Field values assigned in section 4.2.6 of this document are:

- OSPF Domain Identifier: Extended Types 0005, 0105, and 0205.

- OSPF Route Type: Extended Type 0306

- OSPF Router ID: Extended Type 0107

6. Security Considerations

Security considerations which are relevant in general to BGP/MPLS IP VPNS are discussed in [VPN] and [VPN-AS]. We discuss here only those security considerations that are specific to the use of OSPF as the PE/CE protocol.

A single PE may be running OSPF as the IGP of the SP backbone network, as well as running OSPF as the IGP of one or more VPNs. This requires the use of multiple, independent OSPF instances, so that routes are not inadvertently leaked between the backbone and any VPN. The OSPF instances for different VPNs must also be independent OSPF instances, to prevent inadvertent leaking of routes between VPNs.

OSPF provides a number of procedures which allow the OSPF control messages between a PE and a CE to be authenticated. OSPF "cryptographic authentication" SHOULD be used between a PE and a CE. It MUST be implemented on each PE.

In the absence of such authentication, it is possible that the CE might not really belong to the VPN to which the PE assigns it. It may also be possible for an attacker to insert spoofed messages on the PE/CE link, in either direction. Spoofed messages sent to the CE could compromise the routing at the CE's site. Spoofed messages sent to the PE could result in improper VPN routing, or in a denial of service attack on the VPN.

7. Acknowledgments

Major contributions to this work have been made by Derek Yeung and Yakov Rekhter.

Thanks to Ross Callon, Ajay Singhal, Russ Housley, and Alex Zinin for their review and comments.

8. Authors' Addresses

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

E-mail: erosen@cisco.com

Peter Psenak
Parc Pegasus,
De Kleetlaan 6A
1831 Diegem
Belgium

E-mail: ppsenak@cisco.com

Padma Pillay-Esnault
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134

E-mail: ppe@cisco.com

9. Normative References

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

[OSPFv2] "OSPF Version 2", RFC 2328, Moy, J., April 1998

[OSPF-DN] "Using an LSA Options Bit to Prevent Looping in BGP/MPLS IP VPNs", draft-ietf-ospf-2547-dnbit-04.txt, Rosen, Psenak, Pillay-Esnault, March 2004

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

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

10. Informative References

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

[RIP] "RIP Version 2", Malkin, G., RFC 2453, November 1998

[VPN-AS] "Applicability Statement for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4365, E. Rosen, February 2006

11. Full Copyright Statement

Copyright (C) The Internet Society (2006).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an

"AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

12. Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.