Network Working Group
Internet Draft
Intended status: Informational
Expires: August 2007

Rajiv Asati
Cisco Systems

Raymond Zhang
BT

Tom Nadeau
Cisco Systems

Azhar Sayeed
Cisco Systems

February 23, 2007

BGP/MPLS Traffic Blackhole Avoidance
draft-asati-bgp-mpls-blackhole-avoidance-00.txt

Status of this Memo

Copyright Notice

Abstract

In any BGP based MPLS network such as MPLS VPN [RFC4364], an ingress PE router would continue to attract traffic from the CE router by advertising the prefix reachability, even though the Label Switched Path (LSP) from the ingress PE router to the egress PE router may be broken. This causes the VPN traffic to be dropped inside the MPLS VPN network.

This document proposes a framework to make BGP consider the MPLS path availability to the "NEXT_HOP" (i.e. egress PE router) during the BGP bestpath candidate selection process. This document also defines a local database for storing the MPLS path health information for one or more IP prefixes and its interaction with BGP.

Conventions used in this document

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

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].

Table of Contents

1. Introduction

In the current MPLS VPN architecture [RFC4364], a PE router learns the VPNv4 routes from the remote PE routers either over a PE-PE MP-iBGP session or via a PE-RR MP-iBGP session. The remote PE router(s) accepts the VPNv4 routes with the matching import route-targets and advertise them to the CE router(s). The CE router, in turn, is likely to choose the PE router as the next hop to communicate with the relevant remote VPNv4 destinations.

The existing [RFC4364] architecture assumes the ingress PE router to have a working label switched path (LSP) to the egress PE router that advertised the VPNv4 route.

2. Problem Details

The assumption that the ingress PE router always has a working LSP to the egress PE router that advertised the VPNv4 route, may result in the VPN traffic to be dropped or blackholed inside an MPLS VPN network during an LSP failure event. This is because an ingress PE router could qualify a VPNv4 route (learned via an MP-iBGP session) as a valid route, even though the corresponding next hop is no longer MPLS reachable.

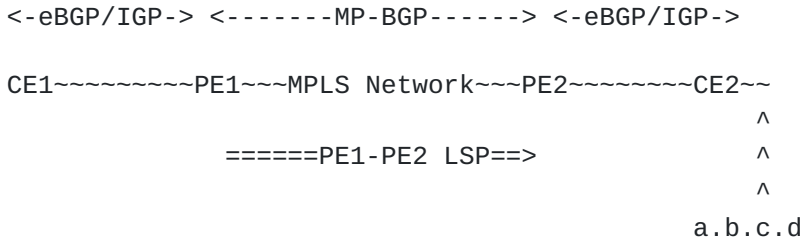

Figure 1 MPLS VPN Network

In the network illustrated in Figure 1, the PE1 to PE2 LSP may be non-functional due to any reason such as down LDP session between the P routers, or the corrupted MPLS Forwarding Table entry, or the missing MPLS Forwarding table entry, or LDP binding defect etc. In such a situation, it is clear that the CE1->CE2 traffic inserted into the MPLS network by PE1 will get dropped inside the MPLS network.

It is undesirable to have PE1 continue to convey to the CE1 router that PE1 (and the MPLS network) is still the next-hop for the remote VPN reachability, without being sure of the corresponding LSP health.

2.1. VPN Deployment Scenarios

It is important to understand the downside of the current framework's limitation using the following two deployment scenarios -

2.1.1. Multi-Homed VPN Site

If the remote VPN site is dual-homed to both PE2 and PE3, then PE1 may learn two VPNv4 paths to the prefix a.b.c.d. via PE2 and PE3 routers, as shown below in Figure 2. PE1 may select the bestpath for the prefix a.b.c.d via PE2 (say, for which the PE1->PE2 LSP is mal-functioning) and advertise that bestpath to CE1 in the context of figure 2.

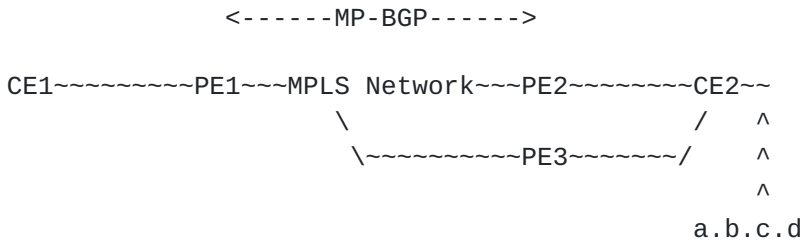

Figure 2 MPLS VPN Network - CE2 Dual-Homing

This causes CE1 to likely send the traffic destined to prefix a.b.c.d to the PE1 router, which forwards the traffic over the malfunctioning LSP to PE2. It is clear that this MPLS encapsulated VPN traffic ends up getting dropped or blackholed somewhere inside the MPLS network.

It is desirable to force PE1 to select an alternate bestpath via that next-hop (such as PE3), whose LSP is correctly functioning.

2.1.2. Single-Homed VPN Site with Site-to-Site Backup Connectivity

The local VPN site may have a backup/dial-up link available at the CE router, but the backup link will not even be activated as long as the CE's routing table continues to point to the PE router as the next-hop (over the MPLS/VPN network).

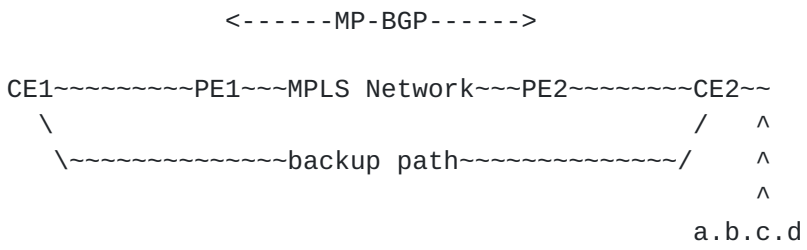

Figure 3 MPLS VPN Network - CE1-CE2 Backup connection

Unless PE2 withdraws the route via the routing protocol used on the PE-CE link, CE1 will not be able to activate the backup link (barring any tracking functionality) to the remote VPN site.

In summary, if PE1 could appropriately qualify the BGP VPNv4 bestpath, then the VPN traffic outage could likely be avoided. Even if the VPN site was not multi-homed, it is desirable to force PE1 to withdraw the path from CE1 to improve the CE-to-CE convergence. This document proposes a mechanism to achieve the optimal BGP behavior at PE.

3. Proposal

The crux of the problem is that the BGP VPNv4 path selection is independent of whether the NEXT_HOP is MPLS reachable or not.

This draft proposes a mechanism to enable BGP to poll whether there is a valid "MPLS path" to the NEXT_HOP of the VPNv4 path, before qualifying that VPNv4 path as the bestpath candidate. This mechanism

comprises of the following three building blocks that are later explained in detail in the subsequent sections.

1. BGP VPNv4 path qualification changes:

    . Qualifies the VPNv4 path as the bestpath candidate only if its NEXT_HOP is MPLS reachable by polling the LSP Health Database.

2. LSP Health Database (LHD):

    . Maintains the information regarding whether a NEXT_HOP is MPLS reachable or not.

3. BGP & LHD Interaction:

    . Specifies the way BGP and LHD interacts with each other.

The proposal helps the ingress PE to either continue to advertise the BGP VPNv4 path's reachability to the CE router by selecting an alternative VPNv4 bestpath, or withdraw the BGP VPNv4 path's reachability from the CE router, during the relevant LSP failure event.

3.1. BGP VPNv4 Path Qualification Changes

As per BGP specification [RFC4271], when a PE router receives a BGP path such as VPNv4 path, BGP qualifies it as the valid candidate for the BGP bestpath using the "Route Resolvability Condition" (Please see section#9.1.2.1 of RFC4271]. Once the path has been qualified as the bestpath candidate, then it gets subjected to the BGP bestpath calculation, which will select the bestpath out of all "bestpath candidates".

The BGP path qualification check-list is highlighted below -

1) NEXT_HOP must be IP reachable

2) ... <any other implementation specific check>

The first check (above) requires the NEXT_HOP of the BGP path to be IP reachable. To determine this reachability, BGP checks the global routing table to validate whether the routing table has a specific or less-specific or default route to the NEXT_HOP. This mechanism is

also referred to as the "NEXT_HOP Validation". The "NEXT_HOP validation" is done not only for the first time when the BGP (VPNv4) path is first received, but also periodically.

The first building block of the proposal is to add another criterion to the "BGP bestpath candidate qualification". The new criterion checks for 'MPLS reachability' to the NEXT_HOP of the BGP path such as VPNv4 path. This means that the NEXT_HOP of the VPNv4 path must be reachable over an MPLS Label Switched Path (LSP). It is desirable to apply this criterion to MPLS only path, as explained in Section 3.1.2.

Hence, with the proposed addition, the path qualification check-list now expands to -

```
   1) NEXT_HOP must be IP reachable

+  2) NEXT_HOP much be MPLS reachable

   3) ... <any other implementation specific check>
```

With the above checks in place, if the NEXT_HOP of a VPNv4 path is not IP reachable, then the path will get marked with the "NEXT_HOP IP Unreachable".

However, if the NEXT_HOP is IP reachable, but not MPLS reachable, then the BGP VPNv4 path will get marked with the "NEXT_HOP MPLS Unreachable". As a result, the BGP VPNv4 path will get disqualified from becoming the bestpath candidate and will not be considered during the bestpath calculation unless the NEXT_HOP becomes MPLS reachable again.

The 'MPLS reachability' to a NEXT_HOP is determined by retrieving the NEXT_HOP specific information from the LSP Health Database (LHD), which is described in section 3.2. The machinery involving BGP and LHD interaction is explained in section 3.3.

3.1.1. IP reachability and MPLS reachability Checks

The BGP path qualification (involving the NEXT_HOP reachability) is performed not only when the path is received for the first time, but also later on using either a timer-driven model or event-driven model or both. This machinery is not modified by the change proposed in

section 3.1. Specifically, the proposal expands the NEXT_HOP reachability check to include checking both:

(1) Routing table aka RIB for the IP reachability, and

(2) LSP Health Database aka LHD for the MPLS reachability.

It is important to note that the LHD check#2 doesn't replace the RIB check#1, but rather complements it. This document doesn't suggest any changes to check#1 whatsoever and assumes that the check#1 will always be performed. Check#2, on the other hand, SHOULD be performed only when appropriate as clarified in section 3.1.2.

One benefit of performing both checks, when appropriate as clarified in section 3.1.2, is in the area of troubleshooting, since it will be clear whether the BGP NEXT_HOP is "IP Unreachable" or "MPLS Unreachable".

3.1.2. 2547oIP based BGP VPNv4 paths

2547oIP technology such as 2547oGRE [RFC4023], 2547oL2TPv3 [RFC3931] etc. doesn't require the usage of the MPLS transport between ingress PE router and egress PE router, hence, it is not desirable to perform the proposed 'MPLS reachability' NEXT_HOP check (check#2) during the BGP VPNv4 path qualification for such BGP paths that utilize IP transport.

> In the simplest form, the above could be achieved by providing a user configurable parameter to enable or disable the check#2 on the router. However, such approach may not work in the deployment involving both 2547oMPLS and 2547oIP BGP VPNv4 paths on the same PE router. Two scenarios in which such deployment may be apparent are (a) the network migration from MPLS to IP or vice versa, (b) the part of network inability to do MPLS.

This section explains a method by which BGP can decide whether to perform the 'MPLS reachability' check (check#2) for a given BGP path.

In this method, the BGP VPNv4 path is flagged (in the relevant BGP data structure) to denote whether BGP VPNv4 path should be subjected to the "'MPLS reachability' to the NEXT_HOP check" during the BGP path qualification. The flag can be updated based on whether the NEXT_HOP information is also conveyed via a separate discovery

mechanism such as a separate BGP AFI/SAFI as defined by existing proposals such as [TUN-SAFI], [BGP-TUN], and forthcoming proposals*.

If the flag is set to one, then the BGP VPNv4 path is considered to belong to 2547oIP, hence, 'MPLS reachability' check is skipped for the NEXT_HOP(s) of such BGP VPNv4 path(s).

If the flag value is zero, the BGP VPNv4 path is considered to belong to 2547oMPLS, hence, 'MPLS reachability' NEXT_HOP check is performed for the NEXT_HOP(s) of such BGP VPNv4 path(s).

This method is advantageous since it appropriately takes care of the deployment scenario in which both 2547oMPLS and 2547oIP based VPNv4 paths may exist on the same PE router.

(* Please note that the forthcoming proposal(s) may let a BGP speaker convey the choice of encapsulation such as MPLS, GRE, L2TPv3 etc. for a given BGP VPNv4 prefix to another BGP speaker. Such proposals are likely to ease the mean of updating the flag discussed here.)

3.2. LSP Health Database (LHD)

As explained in section 3 and 3.1 earlier, BGP will now be required to check for the MPLS reachability to the NEXT_HOP of the BGP VPNv4 path. This means that BGP must somehow obtain the information about NEXT_HOP's MPLS reachability, which is really a forwarding plane element.

Since MPLS reachability relies on the forwarding plane, it is optimal to build a database that would keep the information about whether a NEXT_HOP is reachable over an MPLS LSP or not. This database is referred to as LSP Health Database (LHD).

BGP would use the information from LHD to perform its "NEXT_HOP reachability" check for MPLS reachability. BGP and LHD interaction could be either timer-driven or event-driven depending upon the implementation, though the document may favor the event-driven interaction for faster convergence.

How the LHD is populated is outside the scope of this document and will be covered in a separate document since it may be utilized by other application beyond BGP.

> In short, the LHD could be populated by any 'LSP-health-probe' mechanism such as LSP pings [RFC4389] or BFD LSP [MPLS-BFD] or so forth, to verify whether the LSP to the NEXT_HOP is established or broken.

Although the document expects BGP at an edge router such as PE to utilize the LHD, any other application at any router could utilize the LHD. Three questions become apparent when BGP at PE router needs to utilize the information from the LSP Health Database (LHD) -

1. How would the LHD determine the list of BGP NEXT_HOP(s) that need MPLS reachability check?

2. How frequently should the PE router check the LSP Health for each NEXT_HOP?

3. What actions should BGP take when an LSP starts malfunctioning as recorded in the LHD?

There are at least two ways to figure out the answer to Q1. Please see the next section 3.3 "BGP and LHD interaction" for the detailed answer. For now, let's assume that LHD has the list of NEXT_HOPs i.e. IP addresses to monitor the LSP health for.

Equipped with the list of NEXT_HOPs, the PE router utilizes the 'LSP-health-probe' such as LSP ping, BFD etc. for each NEXT_HOP to validate the LSP health and record it in the LHD. A simplistic view of LHD is shown below -

LSP Health Database Sample::

NEXT_HOP Prefix	LSP Established
192.0.2.11/32	Yes
192.0.2.12/32	Yes
192.0.2.13/32	Yes
192.0.2.14/32	No

Figure 4 LSP Health Database - Simplistic View

Although Q2 is considered specific to the LHD and beyond the scope of this document, it is likely that the PE router may employ either a timer-driven model or event-driven model to update the LHD entries. The frequency of updating the LHD can also be dictated by a user configurable parameter. The frequency should not affect the BGP-LHD interaction machinery.

About Q3, if one or more LSP Health Database (LHD) entries are declared or updated to be broken (shown via "No" in the above simplistic LHD view), then BGP may be notified about it depending on the timer-driven or event-driven model in place. As soon as BGP becomes aware of it, BGP may perform the bestpath calculation i.e. 'BGP VPNv4 path qualification' to explore the alternative bestpaths as discussed in section 3.1. Please see more details about this in next section 3.3.

3.3. BGP and LHD Interaction

At the high level, LSP Health Database (LHD) maintains the LSP health information for one or more addresses that BGP considers as the NEXT_HOPs. BGP utilizes the information from the LHD to declare the 'MPLS reachability' for a NEXT_HOP during the BGP VPNv4 path qualification.

> LHD is constructed and populated using mechanisms that are beyond the scope of this document and will be covered in a different document.

If the LHD entry shows the LSP for a NEXT_HOP to be "established", then BGP will declare the 'MPLS reachability' to the NEXT_HOP check to have passed and rest of the BGP bestpath calculation may continue as usual.

If the LHD shows the LSP for a NEXT_HOP to be not established, then BGP will declare the NEXT_HOP 'MPLS reachability' to have failed and will mark the dependent BGP VPNv4 paths as 'NEXT_HOP MPLS Unreachable'. This will result in such BGP VPNv4 paths be disqualified from becoming the bestpath candidate, and subsequently, PE could update the CE neighbors through one of the following two actions -

> Assuming the presence of alternative BGP VPNv4 path, PE could select a new bestpath, and advertise it to the CE neighbors.

Assuming no other alternative BGP VPNv4 path, PE will withdraw the VPNv4 path(s) from the CE neighbors (independent of the PE-CE routing protocol).

It is obvious that BGP will have to interact with LHD for each of the NEXT_HOPs of the BGP VPNv4 paths. The following logical question then becomes apparent - How does the router determine which LSPs i.e. NEXT_HOPs should be validated within the LHD ?

The document discusses the following two approaches to help LHD obtain the list of NEXT_HOPs -

4. The simplest way is to have the router issue 'LSP-health-probe' for all the host routes since all the NEXT_HOPs are almost always available as the host routes i.e. /32 routes in the routing table. However, every host route is not expected to be the NEXT_HOP, hence, this approach is NOT optimal or accurate since LSP health would be measure for unwanted host routes for no benefits. Moreover, there might be a few fake host routes i.e. /32 routes (including PPP generated /32 routes) that are not really the NEXT_HOPs.

5. The optimal approach is to have the LHD rely on the BGP to provide the list of NEXT_HOPs. BGP should already have a list of the BGP NEXT_HOPs to use it to perform the IP rechability checks. It is logical and appropriate to have LHD perform the LSP Health checks for the NEXT_HOPs specified in this list. Additionally, the list must specify whether LHD should consider all or a subset of the list (since one or more NEXT_HOP may not require MPLS reachability check; This may also depend on the AFI/SAFI of the BGP route). Such inclusion may help to return an appropriate diagnostic code in the MPLS OAM messages such as MPLS ping etc.

Although BGP bestpath calculation and LHD check are independent of each other, one may trigger the other i.e. the BGP bestpath calculation may trigger the LHD check for one or more LHD entries, as well as an LHD entry update may trigger the BGP bestpath calculation for the BGP prefixes learned from the NEXT_HOP pertaining to the LHD entry.

In the case of LHD entry update providing the trigger to perform the BGP bestpath calculation it is desirable to perform the BGP bestpath calculation only for those BGP prefixes whose NEXT_HOP got updated in the LHD to attain an optimal behavior.

LHD will also have to keep up with the changes in the NEXT_HOP list. In other words, if the PE router learns one or more VPNv4 prefix with

the new NEXT_HOP (this could happen when a new PE router is added to the network), then the BGP will update the NEXT_HOP list, which then should be provided to LHD for further processing.

In summary, BGP utilizes the information from the LHD to declare the 'MPLS reachability' to a NEXT_HOP during the BGP VPNv4 path qualification. Although BGP bestpath calculation and LHD check are independent of each other, one may trigger the other.

4. IGP and LHD

This proposal requires neither any changes in the IGP, nor any interaction between IGP and LHD.

5. Applicability

This proposal, although targeted to VPN prefix, can very well be extended to any BGP prefix whose NEXT_HOP utilizes MPLS transport. Few examples are VPNv6, labeled BGP IPv4 prefix [RFC3107], labeled BGP IPv6 prefix etc.

6. Security Considerations

This draft doesn't impose any additional security constraints.

7. IANA Considerations

None.

8. Conclusions

None.

9. Acknowledgments

The authors would like to thank Russ White, Luca Martini, John Monaghan, Chip Popoviciu, Vijay Bollapragada, Carlos Pignataro etc. for their comments and suggestions.

This document was prepared using 2-Word-v2.0.template.dot.

10. References

10.1. Normative References

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

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

[RFC4023] Rosen et al., "Encapsulating MPLS in IP or Generic Routing Encapsulation", RFC4023, March 2005

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

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

[RFC4389] Kompella, K., and Swallo, G., "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006

[RFC3107] Rosen E. and Rekhter Y., "Carrying Label information in BGP-4", RFC 3107, May 2001

10.2. Informative References

[MPLS-BFD] Aggarwal, R., Kompella, K., Nadeau, T., Swallow, G., "BFD for MPLS LSPs", draft-ietf-bfd-mpls, work in progress.

[TUN-SAFI] Nalawade et al, "BGP Tunnel SAFI", draft-nalawade-kapoor-tunnel-safi-05.txt.

[BGP-TUN] Kapoor R., Nalawade G., "BGP4 Tunnel Encapsulation attribute", draft-nalawade-kapoor-idr-bgp-ssa-03.txt, work in progress.

Author's Addresses

Rajiv Asati (Editor)
Cisco Systems
7025 Kit Creek Road
RTP, NC 27560 USA
Email: rajiva@cisco.com

Raymond Zhang
BT
2160 E. Grand Ave.
El Segundo, CA 90245 USA
Email: Raymond_zhang@bt.infonet.com

Tom Nadeau
Cisco Systems
300 Beaver Brook Road
Boxborough, MA, 01719 USA
Email: tnadeau@cisco.com

Azhar Sayeed
Cisco Systems
300 Beaver Brook Road
Boxborough, MA, 01719 USA
Email: asayeed@cisco.com

Intellectual Property Statement

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.

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.