IDR Working Group
Internet Draft
Intended status: Informational
Expires: March 2009

Rajiv Asati
Cisco Systems

October 27, 2008

BGP Bestpath Selection Criteria
draft-asati-idr-bgp-bestpath-selection-criteria-00.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

This Internet-Draft will expire on April 27, 2009.

Copyright Notice

Copyright (C) The IETF Trust (2008).

Abstract

BGP specification [RFC4271] prescribes 'BGP next-hop reachability' as one of the key 'Route Resolvability Condition' that must be satisfied before the BGP bestpath candidate selection. This condition, however, may not be sufficient (as explained in the Appendix section) and desire further granularity.

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

As per BGP specification [RFC4271], when a router receives a BGP path, BGP must qualify it as the valid candidate prior to the BGP bestpath selection using the 'Route Resolvability Condition' (section#9.1.2.1 of RFC4271]. After the path gets qualified as the bestpath candidate, it becomes eligible to be the bestpath, and may get advertised out to the neigbhor(s), if it became the bestpath.

However, in BGP networks that utilize data plane protocol other than IP, such as MPLS etc. to forward the received traffic towards the

next-hop, the above qualification condition may not be sufficient. In fact, this may expose the BGP networks to experience traffic blackholing i.e. traffic loss, due to malfunctioning of the chosen data plane protocol to the next-hop. This is explained further in the Appendix section.

This document defines further granularity to the "Route Resolvability Condition" by (a) resolving the BGP next-hop reachability in the forwarding database of a particular data plane protocol, and (b) optionally including the BGP next-hop "path availability" check.

2. Route Resolvability Condition - Modification

This document proposes two amendments to 'Route Resolvability Condition', which is defined in RFC4271, in consideration for a particular data plane protocol.

1) The next-hop reachability SHOULD be resolved in a particular data plane protocol.

For example, if a BGP IPv4/v6 or VPNv4/v6 path wants to use MPLS data plane to the next-hop, as determined by the policy, then the BGP 'next-hop reachability' should be resolved using the MPLS data plane. In another example, if BGP path wants to use the IP data plane to the next-hop, as determined by the policy, then BGP 'next-hop reachability' should be resolved using the IP data plane. The latter example covers MPLS-in-IP encapsulation techniques such as [RFC4817], [RFC4023] etc.

The selection of particular data plane is a matter of a policy, and is outside the scope of this document. A dynamic signaling such as draft-ietf-softwire-encaps-safi [ENCAP] may be used to convey the data plane protocol chosen by the policy.

This check may be limited to confirming the availability of the valid forwarding entry for the next-hop in the forwarding database of the chosen data plane protocol.

2) The 'path availability' check for the BGP next-hop MAY be performed. This criterion checks for the functioning path to the next-hop in a particular data plane protocol.

The path availability check may be performed by any of the OAM data-plane liveness mechanisms associated with the data plane that is used to reach the Next Hop. The data plane protocol for this criterion must be the same as the one selected by the previous criterion (#1).

The mechanism(s) to perform the "path availability" check and the selection of particular data plane are a matter of a policy and outside the scope of this document.

For example, if a BGP VPNv4 path wants to use the MPLS as the data plane protocol to the next-hop, then MPLS path availability to the next-hop should be evaluated i.e. liveness of MPLS LSP to the next-hop should be validated.

3. Conclusions

Both amendments discussed in section 2 provide further clarity and granularity to help the BGP speaker to either continue to advertise a BGP path's reachability or withdraw the BGP path's reachability, based on the consideration for the path's next-hop reachability and/or availability in a particular data plane.

4. Security Considerations

This draft doesn't impose any additional security constraints.

5. IANA Considerations

None.

6. Acknowledgments

Yakov Rekhter provided critical suggestions and feedback to improve this document. Thanks to John Scudder and Chandrashekhar Appanna for contributing to the discussions that formed the basis of this document. Thanks to Ilya Varlashkin, who made the case to revive this document and provided useful feedback.

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

7. Appendix

7.1. Problem Applicability

In IP networks using BGP, a router would continue to attract traffic by advertising the BGP prefix reachability to neighbor(s) as long as the router had a route to the next-hop in its routing table, but independent of whether the router has a functional forwarding path to the next-hop. This may cause the forwarded traffic to be dropped inside the IP network.

In MPLS or MPLS VPN networks [RFC4364], the same problem is observed if the functional MPLS LSP to the next-hop is not available (due to the forwarding path error on any node along the path to the next-hop).

The following MPLS/VPN topology clarifies the problem -

```
<-eBGP/IGP-> <-------MP-BGP------> <-eBGP/IGP->

CE1~~~~~~~~~PE1~~~MPLS Network~~~PE2~~~~~~~~CE2~~
                                                ^
              ======PE1-PE2 LSP==>              ^
                                                ^
                                            a.b.c.d
```

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 corrupted MPLS Forwarding Table entry, or the missing MPLS Forwarding table entry, or LDP binding defect, or down LDP session between the P routers (with independent label distribution control) 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.

7.1.1. 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.

```
                      <------MP-BGP------>

     CE1~~~~~~~~~PE1~~~MPLS Network~~~PE2~~~~~~~~CE2~~
                          \                      /   ^
                           \~~~~~~~~~~PE3~~~~~~~/    ^
                                                     ^
                                                 a.b.c.d
```

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.

7.1.2. 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).

```
                      <------MP-BGP------>

        CE1~~~~~~~~~PE1~~~MPLS Network~~~PE2~~~~~~~~CE2~~
          \                                         /   ^
           \~~~~~~~~~~~~~~backup path~~~~~~~~~~~~~~/    ^
                                                        ^
                                                    a.b.c.d
```

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.

8. References

8.1. 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.

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

8.2. Informative References

[ENCAP] Rosen, E., Mohapatra, P., "BGP Encapsulation SAFI and BGP Tunnel Encapsulation Attribute", draft-ietf-softwire-encaps-safi-03.txt, work in progress.

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

[RFC4817] Townsley, et al., "Encapsulation of MPLS over Layer 2 Tunneling Protocol Version 3", RFC4817, Nov 2006.

Author's Addresses

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

Intellectual Property Statement

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.

Disclaimer of Validity

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, THE IETF TRUST 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.

Copyright Statement

Copyright (C) The IETF Trust (2008).

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.

Acknowledgment

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