GROW Working Group
Internet-Draft
Intended status: Informational

B. Decraene
France Telecom
P. Francois
UCL
C. Pelsser
IIJ
Z. Ahmad
Orange Business Services
A. J. Elizondo Armengol
Telefonica I+D
T. Takeda
NTT
October 23, 2009

Requirements for the graceful shutdown of BGP sessions
draft-ietf-grow-bgp-graceful-shutdown-requirements-01.txt

Status of this Memo

This Internet-Draft will expire on April 22, 2010.

Requirements for the graceful shutdown of BGP sessions

Copyright Notice

Abstract

The BGP protocol is heavily used in Service Provider networks both for Internet and BGP/MPLS VPN services. For resiliency purposes, redundant routers and BGP sessions can be deployed to reduce the consequences of an AS Border Router or BGP session breakdown on customers' or peers' traffic. However simply taking down or even up a BGP session for maintenance purposes may still induce connectivity losses during the BGP convergence. This is no more satisfactory for new applications (e.g. voice over IP, on line gaming, VPN). Therefore, a solution is required for the graceful shutdown of a (set of) BGP session(s) in order to limit the amount of traffic loss during a planned shutdown. This document expresses requirements for such a solution.

Table of Contents

Requirements for the graceful shutdown of BGP sessions

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 RFC 2119.

2. Introduction

The BGP protocol is heavily used in Service Provider networks both for Internet and BGP/MPLS VPN services. For resiliency purposes, redundant routers and BGP sessions can be deployed to reduce the consequences of an AS Border Router or BGP session breakdown on customers' or peers' traffic.

We place ourselves in the context where a Service Provider needs to shut down one or multiple BGP peering link(s) or a whole ASBR. If an alternate path is available, the requirement is to avoid or reduce customer or peer traffic loss during the BGP convergence. Indeed, as an alternate path is available in the Autonomous System (AS), it should be made possible to reroute the customer or peer traffic on the backup path before the BGP session(s) is/are torn down and the forwarding is interrupted on the nominal path.

The requirements also covers the subsequent re-establishment of the BGP session as even this "UP" case can currently trigger route loss and thus traffic loss at some routers.

Currently, the [BGP] and [MP-BGP] do not include any operation to reduce or prevent traffic loss in case of planned maintenance requiring the shutdown of a forwarding resource. When a BGP session is taken down, BGP behaves as if it was a sudden link or router failure. Besides, the introduction of Route Reflectors as per [BGP RR] to solve scalability issues bound to iBGP full-meshes has worsened the duration of routing convergence: some route reflectors may hide the back up path and depending on RR topology more iBGP hops may be involved in the iBGP convergence. On the other hand, some protocols are already considering such graceful shutdown procedure (e.g. [GMPLS G-Shut]).

Note that these planned maintenance operations cannot be addressed by Graceful Restart extensions [BGP GR] as GR only applies when the forwarding is preserved during the control plane restart. On the contrary, Graceful Shutdown applies when the forwarding is interrupted.
A successful approach of such mechanism should minimize the loss of traffic in most foreseen maintenance situations.

3. Problem statement

As per [BGP], when one (or many) BGP session(s) are shut down to perform a link or router maintenance operation, a BGP NOTIFICATION message is sent to the peer and the session is then closed. A protocol convergence is then triggered both by the local router and by the peer. Alternate paths to the destination are selected, if known. If those alternates paths are not known prior to the BGP session shutdown, additional BGP convergence steps are required in each AS to search for an alternate path.

This behavior is not satisfactory in a maintenance situation because the traffic that was directed towards the removed next-hops may be lost until the end of the BGP convergence. As it is a planned operation, a make before break solution should be made possible.

As maintenance operations are frequent in large networks [Reliability], the global availability of the network is significantly impaired by this BGP maintenance issue.

3.1. Example of undesirable BGP routing behavior

To illustrate these problems, let us consider the following example where one customer router "CUST" is dual-attached to two SP routers "ASBR1" and "ASBR2".
ASBR1 and ASBR2 are in the same AS and owned by the same service provider. Both are iBGP client of the route reflector R1.

```
                                '
                          AS1   '      AS2
                                '

                          /-----------ASBR1---
                         /                     \
                        /                       \
                    CUST                         R1
                        \                       /
                 Z/z     \                     /
                          \-----------ASBR2---

                                '
                          AS1   '      AS2
                                '
```

Before the maintenance, packets for destination Z/z use the CUST-ASBR1 link because R1 selects ASBR1's route based on the IGP cost.

Let's assume the service provider wants to shutdown the ASBR1-CUST link for maintenance purposes. Currently, when the shutdown is performed on ASBR1, the following steps are performed:

1. ASBR1 sends a withdraw to its route reflector R1 for the prefix Z/z.
2. R1 runs its decision process, selects the route from ASBR2 and

advertises the new path to ASBR1.

3. ASBR1 runs its decision process and recovers the reachability of Z/z.

Traffic is lost between step 1 when ASBR1 looses its route and step 3 when it discovers a new path.

Note that this is a simplified description for illustrative purpose. In a bigger AS, multiple steps of BGP convergence may be required to find and select the best alternate path (e.g. ASBR1 is chosen based on a higher local pref, hierarchical route reflectors are used...). When multiple BGP routers are involved and plenty of prefixes are affected, the recovery process can take longer than applications requirements.

3.2. Causes of packet loss

The loss of packets during the maintenance has two main causes:
- lack of an alternate path on some routers
- transient routing inconsistency.

Some routers may lack an alternate path because another router is hiding the backup path. This router can be a route reflector only propagating the best path. Or the backup ASBR does not advertise the backup path because it prefers the nominal path. This lack of knowledge of the alternate path is the first target of this requirement draft.

Transient routing inconsistencies happen during iBGP convergence because all routers are not updating their RIB at the same time. This can lead to forwarding loops and then packet drops. This can be avoided by performing only one IP lookup on BGP routes in each AS and by using tunnels (e.g. MPLS LSP) to send packets between ASBRs.

4. Terminology

g-shut initiator: the router on which the session(s) shutdown is (are) performed for the maintenance.

g-shut neighbor: a router that peers with the g-shut initiator via (one of) the session(s) undergoing maintenance.

Affected prefixes: a prefix initially reached via the peering link(s) undergoing maintenance.

Affected router: a router reaching an affected prefix via a peering link undergoing maintenance.

Initiator AS: the autonomous system of the g-shut initiator router.

Neighbor AS(es): the autonomous system(s) of the g-shut neighbor router(s).

5. Goals and requirements

When a BGP session of the router under maintenance is shut down, the router removes the routes and then triggers the BGP convergence on its BGP peers. The goal of BGP graceful shutdown is to initiate the BGP convergence to find the alternate paths before the nominal paths are removed. As a result, before the nominal BGP session is shut down, all routers learn and use the alternate paths. Then the nominal BGP session can be shut down.

As a result, provided an alternate path is available in the AS, the packets are rerouted before the BGP session termination and fewer packets (possibly none) are lost during the BGP convergence process since at any time, all routers have a valid path.

Another goal is to minimize packet loss when the BGP session is re-established following the maintenance.

From the above goals we can derive the following requirements:

a)   A mechanism to advertise the maintenance action to all affected routers is REQUIRED. Such mechanism may be either implicit or explicit. Note that affected routers can be located both in the local AS and in neighboring ASes.

b)   An Internet wide convergence is OPTIONAL. However if the initiator AS and the neighbor AS(es) have a backup path, they MUST be able to gracefully converge before the nominal path is shut down.

c)   The proposed solution SHOULD be applicable to any kind of BGP sessions (e-BGP, i-BGP, i-BGP route reflector client, e-BGP confederations, e-BGP multi hop, MultiProtocol BGP extension...) and any address family. If a BGP implementation allows closing a sub-set of AFIs carried in a MP-BGP session, this mechanism MAY be applicable to this sub-set of AFIs.

Depending on the session type (eBGP, iBGP...), there may be some variations in the proposed solution in order to fit the requirements.

The following cases should be handled in priority:
- The shutdown of an inter-AS link and therefore the shutdown of an eBGP session.
- The shutdown of an AS Border Router and therefore the shutdown of all its BGP sessions
- The shutdown of a customer access router and all of its BGP sessions. In VPN as per [VPN], this router is called a CE and the use

of others protocols than BGP on the PE-CE access link should also be considered (static routes, RIPv2, OSPF, IS-IS...).

d) The proposed solution SHOULD NOT change the BGP convergence behavior for the ASes exterior to the maintenance process. An incremental deployment on a per AS or per BGP session basis SHOULD be made possible. In case of partial deployment the proposed solution SHOULD incrementally improve the maintenance process. The solution SHOULD bring improvements even when one of the two ASes does not support graceful shutdown. In particular, large Service Providers may not be able to upgrade all of the deployed customer premises access routers (CPE).

e) Redistribution or advertisement of (static) IP routes into BGP SHOULD also be covered.

f) The proposed solution MAY be designed in order to avoid transient forwarding loops. Indeed, forwarding loops increase packet transit delay and may lead to link saturation.

g) The specific procedure SHOULD end when the BGP session is closed. The procedure SHOULD be reverted, either automatically or manually, when the session is re-established. During this reversion procedure -when the session is brought up- the procedure SHOULD also minimize packet loss when the nominal path is installed and used again. In particular, it SHOULD be ensured that the backup path is not removed from the routing tables of the effected nodes before it learns the nominal path. In the end, once the planned maintenance is finished and the shutdown resource becomes available again, the nominal BGP routing MUST be reestablished.

The metrics to evaluate and compare the proposed solutions are, in decreasing order of importance:
- The duration of the remaining loss of connectivity when the BGP session is brought down or up
- The applicability to a wide range of BGP and network topologies, especially those described in section 6;
- The duration of transient forwarding loops;
- The additional load introduced in BGP (eg BGP messages sent to peer routers, peer ASes, the Internet).

6. Reference Topologies

In order to benchmark the proposed solutions, some typical BGP topologies are detailed in this section. The solution drafts should state its applicability for each of these possible topologies.

However, solutions SHOULD be applicable to all possible BGP topologies and not only to these below examples.

6.1. E-BGP topologies

6.1.1. 1 ASBR in AS1 connected to two ASBRs in the neighboring AS2

In this topology we have an asymmetric protection scheme between AS1 and AS2:

- On AS2 side, two different routers are used to connect to AS1.
- On AS1 side, one single router with two BGP sessions is used.

```
                   '
             AS1   '      AS2
                   '
             /----------- ASBR2.1
            /      '
           /       '
        ASBR1.1    '
           \       '
            \      '
             \----------- ASBR2.2
                   '
                   '
         AS1       '      AS2
                   '
```

The requirements of section 5 should be applicable to:

- Maintenance of one of the routers of AS2;
- Maintenance of one link between AS1 and AS2, performed either on an AS1 or AS2 router.

Note that in case of maintenance of the whole router, all its BGP session needs to be shutdown.

6.1.2. 2 ASBRs in AS1 connected to 2 ASBRs in AS2

In this topology we have a symmetric protection scheme between AS1 and AS2: on both sides, two different routers are used to connect AS1 to AS2.

```
                       '
                 AS1   '      AS2
                       '
          ASBR1.1----------- ASBR2.1
                       '
                       '
                       '
                       '
                       '
          ASBR1.2----------- ASBR2.2
                       '
```

```
             AS1       '      AS2
                       '
```

The requirements of section 5 should be applicable to:
- Maintenance of any of the ASBR routers (in AS1 or AS2);
- Maintenance of one link between AS1 and AS2 performed either on an AS1 or AS2 router.

6.1.3. 2 ASBRs in AS2 each connected to two different ASes

In this topology at least three ASes are involved. Depending on which routes are exchanged between these ASes, some protection for some of the traffic may be possible.

```
                    '
              AS1   '      AS2
                    '
       ASBR1.1----------- ASBR2.1
          |         '
          |         '
     '''''|''''''''''
          |         '
          |         '
       ASBR3.1----------- ASBR2.2
                    '
          AS3       '      AS2
```

The requirements of section 5 do not translate as easily as in the two previous topologies because we do not require propagating the maintenance advertisement outside of the two ASes involved in an eBGP session.
For instance if ASBR2.2 requires a maintenance affecting ASBR3.1, then ASBR3.1 will be notified. However we do not require for ASBR1.1 to be notified of the maintenance of the eBGP session between ASBR3.1-ASBR2.2.

6.2. I-BGP topologies

We describe here some frequent i-BGP topologies that SHOULD be supported.
Indeed maintenance of an e-BGP session needs to be propagated within the AS so the solution may depend on the specific i-BGP topology.

6.2.1. iBGP Full-Mesh

In this topology we have a full mesh of iBGP sessions:

```
           P1 ------ P2
           | \     / |
           |  \   /  |
           |   \ /   |     AS1
           |   / \   |
           |  /   \  |
         ASBR1.1---ASBR1.2
            \        /
             \      /
        ''''''\''''/''''''''''''
               \  /      AS2
               ASBR2.1
```

When the session between ASBR1.1 and ASBR2.1 undergoes maintenance, it is required that all i-BGP peers of ASBR1.1 reroute traffic to ASBR1.2 before the session between ASBR1.1 and ASBR2.1 is shut down.

6.2.2. Route Reflector

In this topology, route reflectors are used to limit the number of i-BGP sessions.

```
           P1 RR----- P2 RR
           | \       / |
           |  \     /  |
           |   \   /   |     AS1
           |    \ /    |
           |    / \    |
           |   /   \   |
           |  /     \  |
         ASBR1.1     ASBR1.2
            \          /
             \        /
        ''''''\''''''/''''''''''''
               \    /
                \  /         AS2
               ASBR2.1
```

When the session between ASBR1.1 and ASBR2.1 undergoes maintenance, it is required that all BGP routers of AS1 reroute traffic to ASBR1.2 before the session between ASBR1.1 and ASBR2.1 is shut down.

6.2.3. hierarchical Route Reflector

In this topology, hierarchical route reflectors are used to limit the number of i-BGP sessions.

```
        P1/hRR --------  P2/hRR
           |               |
           |               |
           |               |   AS1
           |               |
           |               |

         P3/RR           P4/RR
           |               |
           |               |
           |               |   AS1
           |               |
           |               |
         ASBR1.1         ASBR1.2
            \             /
             \           /
        ''''''\'''''''''/''''''''''''
               \       /
                \     /        AS2
                 ASBR2.1
```

When the session between ASBR1.1 and ASBR2.1 undergoes maintenance, it is required that all BGP routers of AS1 reroute traffic to ASBR1.2 before the session between ASBR1.1 and ASBR2.1 is shut down.

6.2.4. Confederations

In this topology, a confederation of ASs is used to limit the number of i-BGP sessions. Moreover, RRs may be present in the member ASs of the confederation.
Confederations may be run with different sub-options. Regarding the IGP, each member AS can run its own IGP or they can all share the same IGP. Regarding BGP, local_pref may or may not cross the member AS boundaries.
A solution should support the shutdown of eBGP sessions between member-ASs in the confederation in addition to the shutdown of eBGP sessions between a member-AS and an AS outside of the confederation.

```
        ASBR1C.1 ---------- ASBR1C.2
           |                   |
           |                   |
           |       AS1C        |
           |                   |
           |                   |
        """|"""""""""""""""""""|"""
           |        "          |
         ASBR1A.2   "        ASBR1B.2
           |        "          |
           |        "          |
           |  AS1A  "   AS1B   |             AS1
           |        "          |
           |        "          |
         ASBR1A.1   "         ASBR1B.1
            \       "         /
             \      "        /
        ''''''\'''''''''''''/''''''''''''
               \           /
                \         /                   AS2
                  ASBR2.1
```

In the above figure, member-AS AS1A, AS1B, AS1C belong to a confederation of ASs in AS1. AS1A and AS1B are connected to AS2.

In normal operation, for the traffic toward AS2,
. AS1A sends the traffic directly to AS2 through ASBR1A.1
. AS1B sends the traffic directly to AS2 through ASBR1B.1
. AS1C load balances the traffic between AS1A and AS1B

When the session between ASBR1A.1 and ASBR2.1 undergoes maintenance, it is required that all BGP routers of AS1 reroute traffic to ASBR1B.1 before the session between ASBR1A.1 and ASBR2.1 is shut down.

7. Security Considerations

Security considerations MUST be addressed by the proposed solutions.

One AS SHOULD NOT be able to use the graceful shutdown procedure to selectively influence routing decision in the peer AS (inbound TE) outside the case of the planned maintenance. In the case the proposed solution allows this, the peer AS SHOULD have means to detect such behavior.

8. IANA Considerations

This document has no actions for IANA.

9. References

9.1. Normative References

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

[MP-BGP] T. Bates, R. Chandra, D. Katz, Y. Rekhter
"Multiprotocol Extensions for BGP-4", RFC 4760 January 2007.

[BGP RR] T. Bates, E. Chen, R. Chandra
"BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456 April 2006.

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

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

9.2. Informative References

[GMPLS G-Shut] Z. Ali, J.P. Vasseur, A. Zamfir and J. Newton
"Graceful Shutdown in MPLS and Generalized MPLS Traffic Engineering Networks" September 15, 2009, internet draft, draft-ietf-ccamp-mpls-graceful-shutdown-12.txt, work in progress.

[Reliability] Network Strategy Partners, LLC.
"Reliable IP Nodes: A prerequisite to profitable IP services", November 2002. http://www.nspllc.com/NewPages/Reliable_IP_Nodes.pdf

10. Acknowledgments

This draft is mostly an updated version of draft-dubois-bgp-pm-reqs-02.txt.

Authors would like to thank Nicolas Dubois, Benoit Fondeviole, Christian Jacquenet, Olivier Bonaventure, Steve Uhlig, Xavier Vinet, Vincent Gillet, Jean-Louis le Roux and Pierre Alain Coste for the useful discussions on this subject, their review and comments.

This draft has been partly sponsored by the European project IST AGAVE.

Authors' Addresses

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
92794 Issy Moulineaux cedex 9
France
Email: bruno.decraene@orange-ftgroup.com

Pierre Francois
Universite catholique de Louvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348
BE
Email: francois@info.ucl.ac.be

Cristel Pelsser
Internet Initiative Japan
Jinbocho Mitsui Building
1-105 Kanda jinbo-cho
Chiyoda-ku, Tokyo 101-0051
Japan
Email: cristel@iij.ad.jp

Zubair Ahmad
Orange Business Services
13775 McLearen Road, Oak Hill VA 20171
USA
Email: zubair.ahmad@ orange-ftgroup.com

Antonio Jose Elizondo Armengol
Division de Analisis Tecnologicos
Technology Analysis Division
Telefonica I+D
C/ Emilio Vargas 6
28043, Madrid
E-mail: ajea@tid.es

Tomonori Takeda
NTT Corporation
9-11, Midori-Cho 3 Chrome
Musashino-Shi, Tokyo 180-8585
Japan
Email: takeda.tomonori@lab.ntt.co.jp