Network Working Group
N. Dubois
B. Decraene
B. Fondeviole
France Telecom
Internet Draft
Z. Ahmad
Equant
Document: draft-dubois-bgp-pm-reqs-02.txt
July 2005
Expiration Date: January 2006

Requirements for planned maintenance of BGP sessions

draft-dubois-bgp-pm-reqs-02.txt

Status of this Memo

Abstract

To ease the maintenance of BGP [BGP] sessions and limit the amount of traffic that is lost during planned maintenance operations on routers, a solution is required in order to gracefully shutdown a router or a session.

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 [1].

Table of Contents

1. Introduction

The BGP protocol is heavily used in Service Provider networks. For resiliency purposes, most of the IP network operators deploy redundant routers and BGP sessions to minimize the risk of BGP session breakdown towards their customers or peers.
In a context where a Service Provider wants to upgrade or remove a particular router that maintains one or several BGP sessions, our requirement is to avoid customer or peer traffic loss as much as possible. It should be made possible to reroute the customer or peer traffic before the maintenance operation occurs and BGP session is torn down.

Currently, the BGP specification does not include any operation to prevent traffic loss in case of planned maintenance.

A successful approach of such mechanism should indeed minimize the loss of traffic in most foreseen maintenance situations. It should be easily deployable and if possible, provide backward compatibility.

2. Problem Statement

Currently, when one (or many) BGP session needs to be shut down, a BGP NOTIFICATION message is sent to the peer and the session is then closed. A protocol convergence is then triggered both in the local

router and in the peer. Alternate routes to the destination are selected, if available.

This behavior is not satisfactory in a maintenance situation because customer's (or peerÆs) traffic that was directed towards the removed next-hops is lost until the end of 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, the global availability of the network is significantly impaired by the BGP maintenance issues.

3. Terminology

Maintained router: The router undergoing maintenance, closing (a) BGP session(s) and causing the rerouting.

Peer routers: Routers which have a BGP peering session with the Maintained router.

Impacted routers: Routers which use the Maintained router as a BGP Next Hop.

4. Goals and Requirements

When some or all BGP sessions of a Maintained router need to be administratively shut down, instead of sending a BGP NOTIFICATION message and/or tearing the TCP session down, our goal is to achieve the following behavior:

First problem : session stops

Step 1:
A mechanism is implemented on the Maintained router in order to gracefully reroute packets towards and from the BGP next-hop that is going to be unavailable.
By doing so, packets are rerouted before the maintenance operation and no packet is lost for all the destination prefixes for which an alternate route is available. The proposed solution MAY be designed in order to avoid transient routing loops.

Step 2:
Once traffic is correctly rerouted BGP sessions are shutdown.

Second problem: session starts

Step 3:
Once maintenance operation has been completed, a mechanism may be implemented to gracefully restore traffic to the original path

avoiding transient routing loops.

Summary:

As a result, if another router provides an alternate path towards a set of destination prefixes, the packets are rerouted before the BGP session termination and no packet is lost during BGP convergence process, since both the forwarding and the Loc-RIB tables are kept while the peers are re-computing their forwarding tables.

From the above goals we can derive the following requirements:

a/ A mechanism to advertise the maintenance action to all Impacted routers is REQUIRED. Such mechanism may be either implicit or explicit.
Note that Impacted routers can be located in adjacent ASes. The proposed solution MAY be designed in order to avoid transient routing loops.

b/ It is REQUIRED that the Maintained router implements a mechanism to keep the forwarding for the NLRI undergoing maintenance until all reroutable packets has been rerouted.

c/ A mechanism may be needed to indicate the end of the graceful maintenance operation. The proposed solution MAY be designed in order to avoid transient routing loops.

d/ An Internet wide convergence is NOT REQUIRED. However the local AS and its directly connected peersÆ ASes MUST be able to gracefully converge before the service interruption.

e/ The proposed solution SHOULD be applicable to all kinds of BGP sessions (e-BGP/MP-eBGP, i-BGP/MP-iBGP and i-BGP/MP-iBGP route reflector client) and any address family. Depending on the session type, there may be some variation in the proposed solution in order to fit the requirement. If the BGP implementation allows closing a sub-set of AFIs carried in a MP-BGP session, this mechanism is applicable to this sub-set of AFI identifiers. However the following cases should be handled first:

- The maintenance of one particular e-BGP/MP-eBGP session.
- The reload of one AS border router.
- The shutdown of PE <-> CE links (Static & eBGP) in a MPLS-VPN environment.

f/ 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 basis MUST be made possible. It means that the proposed solution SHOULD be interoperable with the current BGP implementation and SHOULD improve the maintenance process even when one of the two ASes does not support graceful maintenance. In particular, large BGP/MPLS VPN Service Providers may

not be able to upgrade all of the deployed CEs. The solution SHOULD

improve the behavior during planned maintenance even with Vanilla CEs.

g/ If possible, redistribution of static IP routes into iBGP/MP-iBGP SHOULD also be covered. Indeed, static routes are often used between PE and CE in a BGP/MPLS VPN environment.

5. Scope

The purpose of this requirement is neither to solve all the convergence issues that may arise within the Internet nor to modify the convergence properties of the BGP protocol.

The Example section illustrates typical and important cases where this requirement should be applicable and tries to make it more understandable.

In addition a Reference Topologies section presents some BGP topologies (both i-BGP and e-BGP) and confronts them to the requirement. These topologies SHOULD be used to test the proper behavior of proposed solutions.

6. Example

Purpose of this section is to give one typical example. It should help the reader to understand how graceful maintenance will enhance the availability of the inter provider BGP connections.

Let us consider the following example (Figure 1 below) where one customer router (denoted as "CUST" in the figure) is dual-homed to two SP routers, denoted as "ASBR1" and "ASBR2". ASBR1 and ASBR2 are in the same AS and owned by the same service provider.

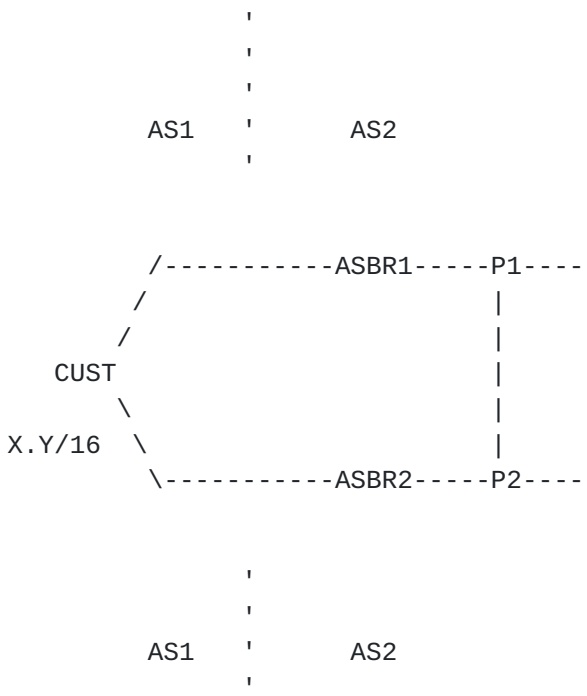


Figure 1: Redundant peering example.

Packets are normally conveyed by the CUST-ASBR1 link. Let's assume the service provider wants to shutdown ASBR1 for maintenance purposes.

The behavior as defined in [BGP] is:
1. ASBR1 tears down all of its BGP sessions.
2. As a result, it removes all the BGP routes from its RIB and FIB tables.
3. Peers routers remove all the routes that were announced by the shutting down peer and advertise the failure to all their BGP peers. These peers are likely Impacted routers.
4. Impacted routers, receive BGP update messages, perform a BGP selection process and update their RIB and FIB accordingly.

During Impacted routers' convergence:

- CUST continues to send packets to ASBR1. ASBR1 drops these packets because it has no route to destination.
- P1 and possibly P2 continue to send traffic to ASBR1. ASBR1 drops this traffic because it has no route to CUST (X.Y/16).

From the customer's point of view, packets are lost during the BGP convergence time.

With the required behavior defined in section 4 [Goals and Requirements]:

- On all of its BGP sessions, ASBR1 signals a maintenance according to the requirement defined in section 4-a.

- During the BGP convergence of all Impacted routers, ASBR1 keeps forwarding customer traffic in both directions.
- Once traffic has been rerouted, ASBR1 closes its BGP sessions with its peers. No packet is lost.

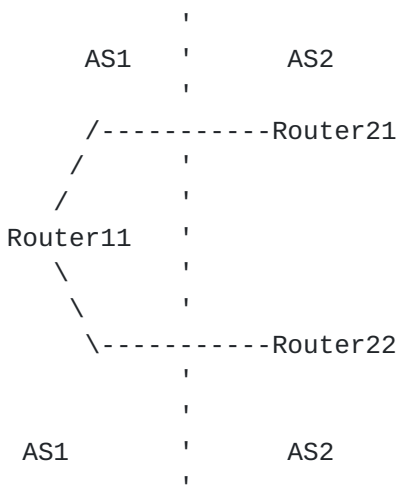
7. Reference Topologies

In order to benchmark the proposed solutions, some typical BGP topologies are detailed. Solutions SHOULD be applicable to all topologies described below.
A solution draft should study the applicability of its solution for each of these 9 (3 E-BGP * 3 I-BGP) possible topologies.

Terminology used in this section is inspired from RFC 2547. We use PE (provider edge router) and CE (customer edge router). However the scope of applicability is broader and can be transposed to any inter ûAS BGP peering solution.

7.1. E-BGP/MP-eBGP topologies

Topology 1CE <-> 2PE:



In this topology we have an asymmetric protection scheme between AS 1 and AS 2:

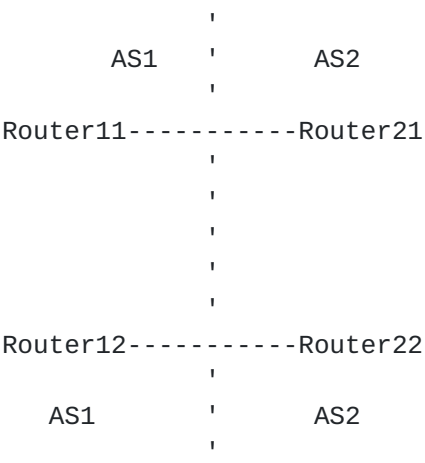
- On AS 2 side, two different routers have been used to connect to AS 1.
- On AS 1 side, one single router with two BGP sessions is used.

The requirement of section 4 should be applicable to:

- Maintenance of one of the routers of AS2.
- Maintenance of the router of AS1.

- Maintenance of one of the two sessions between AS1 and AS2.

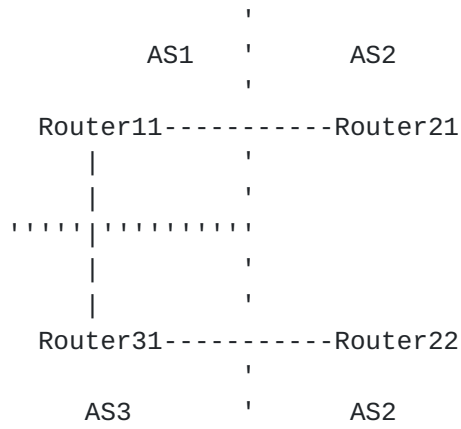
Topology 2CE <-> 2PE:



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

The requirement of section 4 should be applicable to:
- Maintenance of any of the routers (in AS1 or AS2).
- Maintenance of one of the two sessions between AS1 and AS2.

Topology 2CE <-> 2ISP:



In this topology the protection scheme between AS1 and AS2 is not as symmetric as in the two previous topologies. Depending on which routes are exchanged between the 3 ASes, some protection for some of the traffic may be possible.

The requirement of section 4 does not translate as easily as in the two previous topologies because we do not require propagating the maintenance advertisement in the Internet.

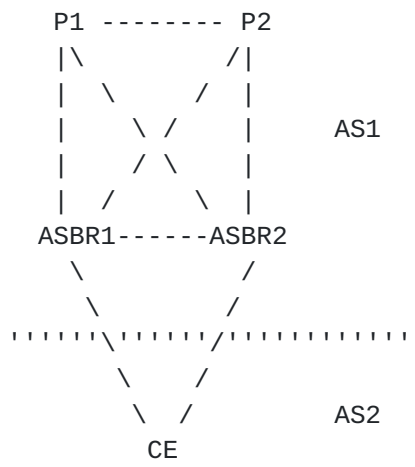
For instance if Router22 requires a maintenance impacting Router31, then Router31 will be notified. However we do not require for Router11 to be notified.

7.2. I-BGP/MP-iBGP topologies

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

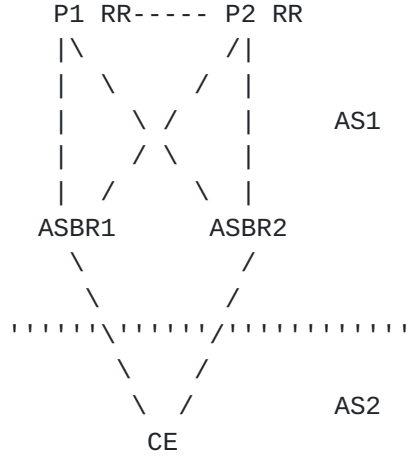
Topology "Full-Mesh":

It is a full iBGP mesh topology as represented below.



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

Topology "RR":

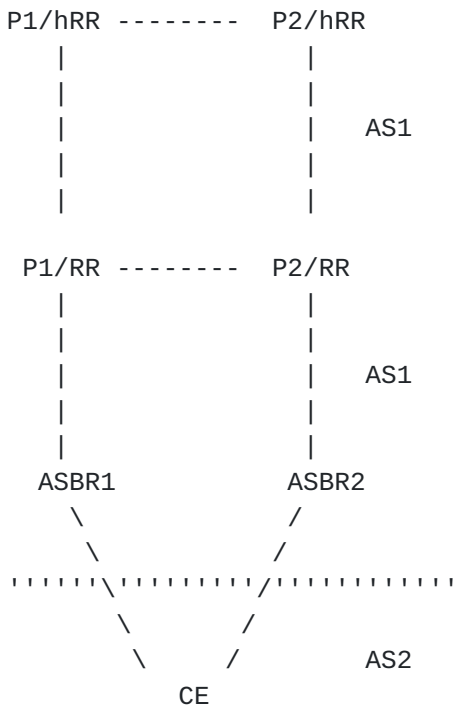


In this topology, route reflectors are used to limit the number of i-BGP sessions.

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

Topology "hierarchical RR":

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



8. Security Considerations

Security consideration MUST be addressed by the proposed solutions.

9. Intellectual Property Considerations

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to per-

tain to the implementation or use of the technology described in this document or the extent to which any license under such rights might

10. Acknowledgments

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

11. References

[BGP] Y. Rekhter, T. Li,
"A Border Gateway protocol 4 (BGP)", RFC 1771, March 1995.

[MP-BGP] T. Bates, Y. Rekhter, R. Chandra, D. Katz,
"Multiprotocol Extensions for BGP-4", RFC 2858 June 2000.

12. Author's Addresses

Nicolas Dubois
France Telecom
24, rue du G n ral Bertrand
75007 Paris
France
Email: nicolas.dubois@francetelecom.com

Bruno Decraene
France Telecom
38-40 rue de general Leclerc
92794 Issy Moulineaux cedex 9
France
Email: bruno.decraene@francetelecom.com

Benoit Fondeviole

France Telecom
38-40 rue de general Leclerc

92794 Issy Moulineaux cedex 9
France
Email: benoit.fondeviole@francetelecom.com

Zubair Ahmad
Equant
13775 McLearen Road, Oak Hill VA 20171
USA
Email: zubair.ahmad@equant.com