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 22, 2010

Requirements for the graceful shutdown of BGP sessions
draft-ietf-grow-bgp-graceful-shutdown-requirements-06.txt

Status of this Memo

Copyright Notice

Abstract

The Border Gateway Protocol(BGP) 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 bringing up a BGP session for maintenance purposes may still induce connectivity losses during the BGP convergence. This is not satisfactory any more 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

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 [RFC2119].

2. Introduction

The Border Gateway Protocol(BGP) [BGP-4] is heavily used in Service Provider networks both for Internet and BGP/MPLS VPN services [VPN]. 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 performs a maintenance operation and needs to shut down one or multiple BGP peering link(s) or a whole ASBR. If an alternate path is available within the AS, 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reroute the customer or peer traffic on this backup path before the BGP session(s) is/are torn down, the nominal path withdrawn and the forwarding is interrupted on the nominal path.

The requirements also cover 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, BGP [BGP-4] and MP-BGP [MP-BGP] do not include any operation to gracefully advertise or withdraw a prefix while traffic toward that prefix could still be correctly forwarded using the old path. When a BGP session is taken down, BGP behaves as if it was a sudden link or router failure and withdraws the prefixes learnt over that session, which may trigger traffic loss. There is no mechanism to advertise to its BGP peers that the prefix will soon be unreachable, while still being reachable. When applicable, such mechanism would reduce or prevent traffic loss. It would typically be applicable in case of a maintenance operation requiring the shutdown of a forwarding resource. Typical examples would be a link or line card maintenance, replacement or upgrade. It may also be applicable for a software upgrade as it may involve a firmware reset on the line cards and hence forwarding interruption.
The introduction of Route Reflectors as per [RR] to solve scalability issues bound to IBGP full-meshes has worsened the duration of routing convergence as some route reflectors may hide the back up path. Thus depending on RR topology more IBGP hops may be involved in the IBGP convergence.

Note that these planned maintenance operations cannot be addressed by Graceful Restart extensions [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.
Note also that some protocols are already considering such graceful shutdown procedure (e.g. GMPLS in [RFC5817]).

A successful approach of such mechanism should minimize the loss of traffic in most foreseen maintenance situations.

3. Problem statement

As per [BGP-4], when one (or many) BGP session(s) are shut down, 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 simple 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.

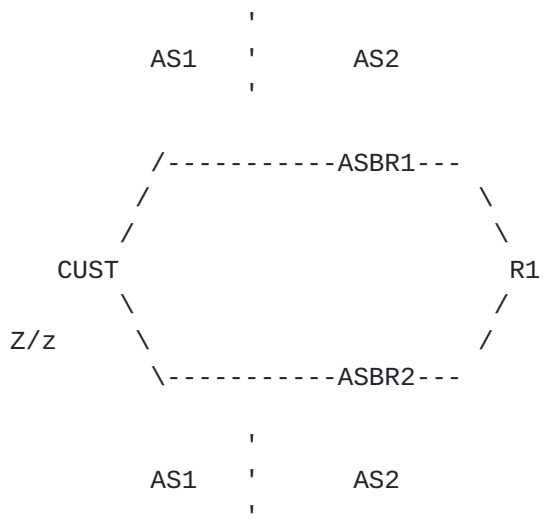

Figure 1. Dual attached customer

Before the maintenance, packets for destination Z/z use the ASBR1-CUST 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 its best path;
- the backup ASBR not advertising 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 and FIB at the same time. This can lead to forwarding loops and then packet drops. The duration of these transient micro-loops may depend on the IBGP topology (e.g. number of Route Reflectors between ingress and egress ASBR), implementation differences among router platforms (e.g. speed to update the RIB and FIB, possibly the order in which prefixes are modified), forwarding mode (hop by hop IP forwarding versus tunneling).
Transient forwarding loops 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. As such, BGP/MPLS VPNs should be immune to such micro forwarding loops.

4. Terminology

g-shut: Graceful SHUTdown. A method for explicitly notifying the BGP routers that a BGP session (and hence the prefixes learnt over that session) is going to be disabled.

g-noshut: Graceful NO SHUTdown. A method for explicitly notifying the BGP routers that a BGP session (and hence the prefixes learnt over that session) is going to be enabled.

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 with enough remaining capacity 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. Note also that the maintenance action can either be the shutdown of a BGP session or the establishment of a BGP session.
The mechanism SHOULD allow BGP routers to minimize packet loss when a path is removed or advertised. In particular, it SHOULD be ensured that the old path is not removed from the routing tables of the affected routers before the new path is known.
The solution mechanism MUST reduce packet loss but MAY provide only a reduction rather than full minimization, in order to trade off with simplicity of implementation and operation as shown in some of the following requirements.

b) An Internet wide convergence is OPTIONAL. However if the initiator AS and the neighbor AS(es) have a backup path, they SHOULD 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 (EBGP, IBGP, IBGP route reflector client, EBGP confederations, EBGP multi hop, MultiProtocol BGP extension...) and any address family. If a BGP implementation allows closing or enabling 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 kind of session, there may be some variations in the proposed solution in order to fulfill 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.

Service Providers and platforms implementing a graceful shutdown solution should note that in BGP/MPLS VPN as per [VPN], the PE-CE routing can be performed by other protocols than BGP (e.g. static routes, RIPv2, OSPF, IS-IS). This is out of scope of this document.

d) The proposed solution SHOULD NOT change the BGP convergence behavior for the ASes exterior to the maintenance process, namely ASes other than the initiator AS and it(s) neighbor AS(es).

e) An incremental deployment on a per AS or per BGP session basis MUST be made possible. In case of partial deployment the proposed solution SHOULD incrementally improve the maintenance process.
It should be noted that in an inter domain relation, one AS may have more incentive to use graceful shutdown than the other. Similarly, in a BGP/MPLS VPN environment, it's much easier to upgrade the PE routers than the CE mainly because there is at least an order of magnitude more CE and CE locations than PE and PE locations. As a consequence, when splitting the cost of the solution between the g-shut initiator and the g-shut neighbour the solution SHOULD favour a low cost solution on the neighbour AS side in order to reduce the impact on the g-shut neighbour. Impact should be understood as a generic term which includes first hardware, then software, then configuration upgrade..

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

g) 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.

h) The specific procedure SHOULD end when the BGP session is closed following the g-shut and once the BGP session is gracefully opened following the g-noshut. In the end, once the planned maintenance is finished the nominal BGP routing MUST be reestablished.
The duration of the g-shut procedure, and hence the time before the BGP session is safely closed SHOULD be discussed by the solution document. Examples of possible solutions are the use of a pre-configured timer, of a message to signal the end of the BGP convergence or monitoring the traffic on the g-shut interface...

i) The solution SHOULD be simple and simple to operate. Hence it MAY only cover a subset of the cases. (As a consequence, most of the

above requirements are expressed as "SHOULD" rather than "MUST")

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 simplicity;
- 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 documents should state the applicability of the solutions for each of these possible topologies.

However, solutions SHOULD be applicable to all possible BGP topologies and not only to these below examples. Note that this is a "SHOULD" rather than a "MUST" as a partial lightweight solution may be preferred to a full but more complex solution. Especially since some ISP may not be concerned by some topologies (e.g. confederations).

6.1. EBGP topologies

We describe here some frequent EBGP topologies that SHOULD be supported by the solution.

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.

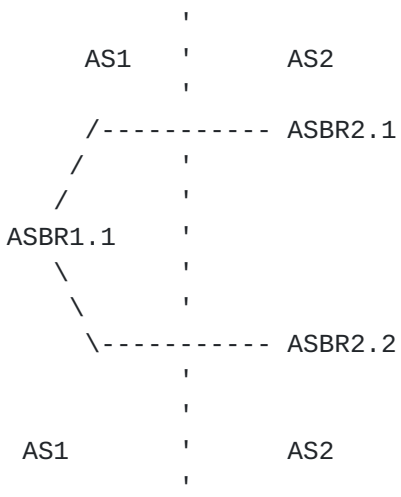

Figure 2. EBGP topology with redundant ASBR in one of the AS.

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 sessions need to be gracefully shutdown at the beginning of the maintenance and gracefully brought up at the end of the maintenance.

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.

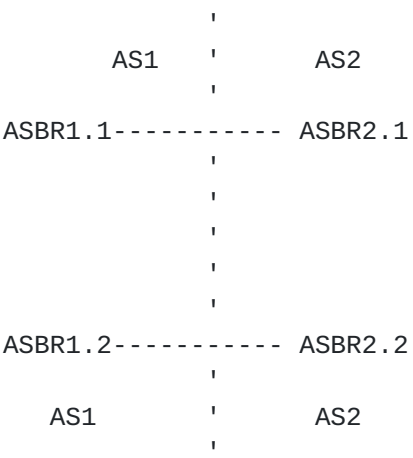

Figure 3. EBGP topology with redundant ASBR in both ASes

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.

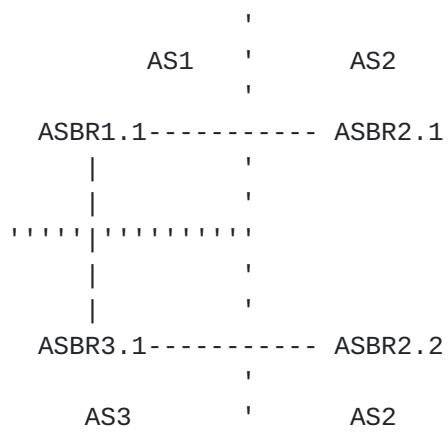

Figure 4. EBGP topology of a dual homed customer

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. IBGP topologies

We describe here some frequent IBGP topologies that SHOULD be supported by the solution.

6.2.1. IBGP Full-Mesh

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

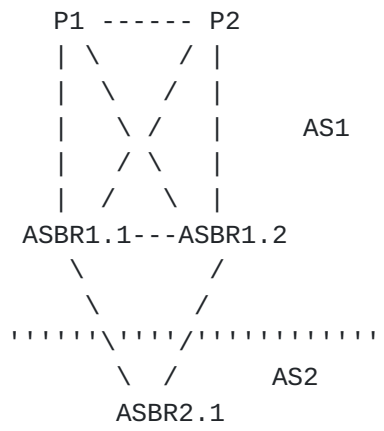

Figure 5. IBGP full mesh

When the session between ASBR1.1 and ASBR2.1 is gracefully shutdown, it is required that all routers of AS1 reroute traffic to ASBR1.2 before the session between ASBR1.1 and ASBR2.1 is shut down.
Symmetrically, when the session between ASBR1.1 and ASBR2.1 is gracefully brought up, it is required that all routers of AS1 preferring ASBR1.1 over ASBR1.2 reroute traffic to ASBR1.1 before the less preferred path trough ASBR1.2 is possibly withdrawn.

6.2.2. Route Reflector

In this topology, route reflectors are used to limit the number of IBGP sessions. There is a single level of route reflectors and the route reflectors are fully meshed.

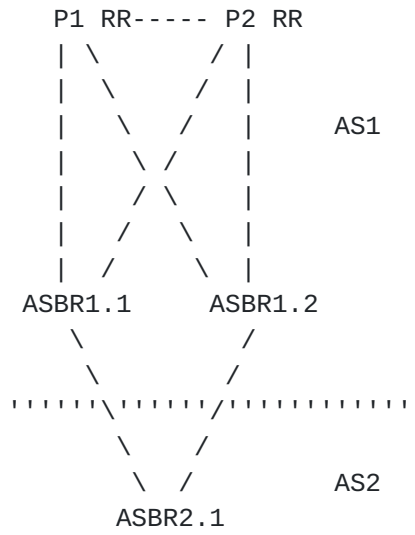

Figure 6. Route Reflector

When the session between ASBR1.1 and ASBR2.1 is gracefully shutdown, 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.
Symmetrically, when the session between ASBR1.1 and ASBR2.1 is gracefully brought up, it is required that all routers of AS1 preferring ASBR1.1 over ASBR1.2 reroute traffic to ASBR1.1 before the less preferred path trough ASBR1.2 is possibly withdrawn.

6.2.3. hierarchical Route Reflector

In this topology, hierarchical route reflectors are used to limit the number of IBGP sessions. There could me more than levels of route reflectors and the top level route reflectors are fully meshed.

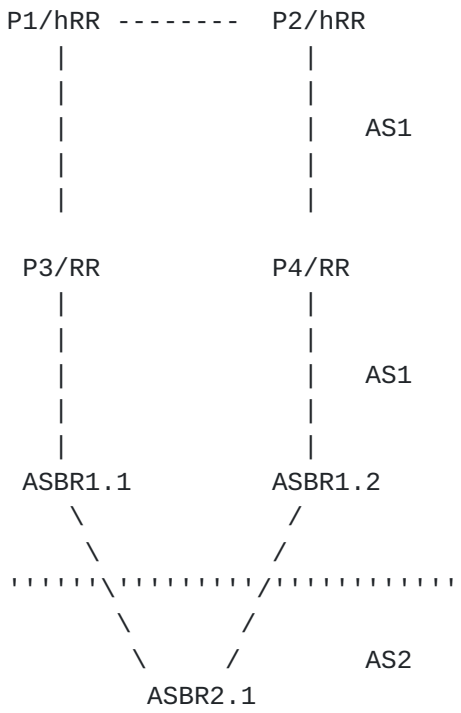

Figure 7. Hierarchical Route Reflector

When the session between ASBR1.1 and ASBR2.1 is gracefully shutdown, 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.
Symmetrically, when the session between ASBR1.1 and ASBR2.1 is gracefully brought up, it is required that all routers of AS1 preferring ASBR1.1 over ASBR1.2 reroute traffic to ASBR1.1 before the less preferred path trough ASBR1.2 is possibly withdrawn.

6.2.4. Confederations

In this topology, a confederation of ASs is used to limit the number of IBGP 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 graceful shutdown and graceful bring up of EBGP sessions between member-ASs in the confederation in addition to the graceful shutdown and graceful bring up of EBGP sessions between a member-AS and an AS outside of the confederation.

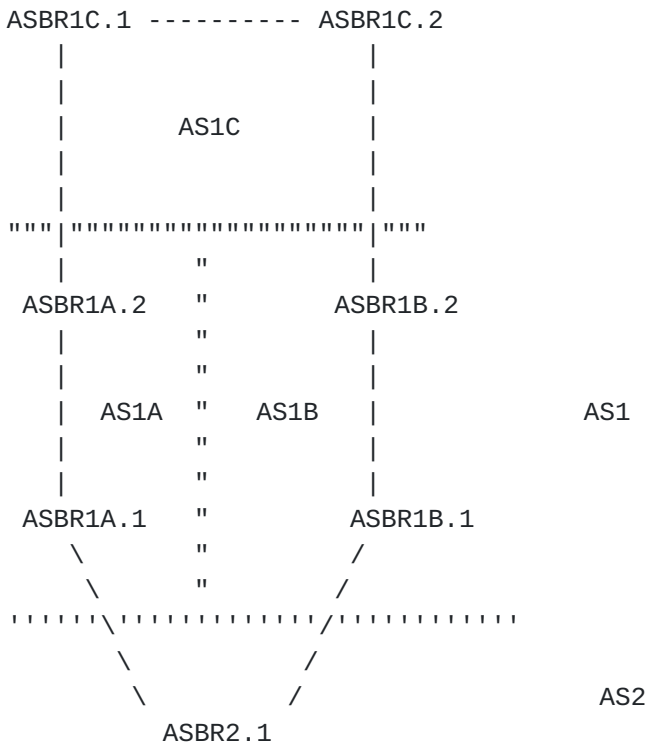

Figure 8. Confederation

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 is gracefully shutdown, 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.
Symmetrically, when the session between ASBR1A.1 and ASBR2.1 is gracefully brought up, it is required that all routers of AS1 preferring ASBR1A.1 over ASBR1.2 reroute traffic to ASBR1A.1 before the less preferred path trough ASBR1.2 is possibly withdrawn.

7. Security Considerations

At the requirements stage, this graceful shutdown mechanism is expected to not affect the security of the BGP protocol, especially if it can be kept simple. No new sessions are required and the additional ability to signal the graceful shutdown is not expected to

bring additional attack vector as BGP neighbors already have the ability to send incorrect or misleading information or even shut down the session.

Security considerations MUST be addressed by the proposed solutions. In particular they SHOULD address the issues of bogus g-shut messages and how they would affect the network(s), as well as the impact of hiding a g-shut message so that g-shut is not performed.

The solution SHOULD NOT increase the ability for one AS to selectively influence routing decision in the peer AS (inbound Traffic Engineering) outside the case of the BGP session shutdown. Otherwise, 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

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

[BGP-4] 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.

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

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

9.2. Informative References

[RFC5817] Z. Ali, J.P. Vasseur, A. Zamfir and J. Newton
"Graceful Shutdown in MPLS and Generalized MPLS Traffic Engineering Networks", RFC 5817 April 2010.

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

[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

Authors would like to thank Nicolas Dubois, Benoit Fondeviole, Christian Jacquenet, Olivier Bonaventure, Steve Uhlig, Xavier Vinet, Vincent Gillet, Jean-Louis le Roux, Pierre Alain Coste and Ronald Bonica 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