Network Working Group
Internet-Draft
Intended status: Informational
Expires: June 10, 2012

Pierre Francois
Institute IMDEA Networks
Bruno Decraene
France Telecom
Cristel Pelsser
Internet Initiative Japan
Keyur Patel
Clarence Filsfils
Cisco Systems
December 8, 2011

Graceful BGP session shutdown
draft-ietf-grow-bgp-gshut-03

Abstract

This draft describes operational procedures aimed at reducing the amount of traffic lost during planned maintenances of routers or links, involving the shutdown of BGP peering sessions.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on June 10, 2012.

Table of Contents

1. Introduction

Routing changes in BGP can be caused by planned, maintenance operations. This document discusses operational procedures to be applied in order to reduce or eliminate losses of packets during the maintenance. These losses come from the transient lack of reachability during the BGP convergence following the shutdown of an eBGP peering session between two Autonomous System Border Routers (ASBR).

This document presents procedures for the cases where the forwarding plane is impacted by the maintenance, hence when the use of Graceful Restart does not apply.

The procedures described in this document can be applied to reduce or avoid packet loss for outbound and inbound traffic flows initially forwarded along the peering link to be shut down. These procedures trigger, in both involved ASes, rerouting to the alternate path, while allowing routers to keep using old paths until alternate ones are learned, installed in the RIB and in the FIB. This ensures that routers always have a valid route available during the convergence process.

The goal of the document is to meet the requirements described in [REQS] at best, without changing the BGP protocol.

Still, it explains why reserving a community value for the purpose of BGP session graceful shutdown would reduce the management overhead bound with the solution. It would also allow vendors to provide an automatic graceful shutdown mechanism that does not require any router reconfiguration at maintenance time.

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. Terminology

g-shut initiator: a router on which the session shutdown is performed for the maintenance.

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

Initiator AS: the Autonomous System of the g-shut initiator.

Neighbor AS: the Autonomous System of the g-shut neighbor.

Loss of Connectivity (LoC: the state when a router has no path towards an affected prefix.

3. Packet loss upon manual eBGP session shutdown

Packets can be lost during a manual shutdown of an eBGP session for two reasons.

First, routers involved in the convergence process can transiently lack of paths towards an affected prefix, and drop traffic destined to this prefix. This is because alternate paths can be hidden by nodes of an AS. This happens when the paths are not selected as best by the ASBR that receive them on an eBGP session, or by Route Reflectors that do not propagate them further in the iBGP topology because they do not select them as best.

Second, within the AS, the FIB of routers can be transiently inconsistent during the BGP convergence and packets towards affected prefixes can loop and be dropped. Note that these loops only happen when ASBR-to-ASBR encapsulation is not used within the AS.

This document only addresses the first reason.

4. Practices to avoid packet losses

This section describes means for an ISP to reduce the transient loss of packets upon a manual shutdown of a BGP session.

4.1. Improving availability of alternate paths

All solutions that increase the availability of alternate BGP paths at routers performing packet lookups in BGP tables such as [BestExternal] and [AddPath] help in reducing the LoC bound with manual shutdown of eBGP sessions.

One of such solutions increasing diversity in such a way that, at any single step of the convergence process following the eBGP session shutdown, a BGP router does not receive a message withdrawing the only path it currently knows for a given NLRI, allows for a simplified g-shut procedure.

Note that the LoC for the inbound traffic of the maintained router, induced by a lack of alternate path propagation within the iBGP topology of a neighboring AS is not under the control of the operator performing the maintenance. The part of the procedure aimed at avoiding LoC for incoming paths can thus be applied even if no LoC

are expected for the outgoing paths.

4.2. Make before break convergence: g-shut

This section describes configurations and actions to be performed to perform a graceful shutdown procedure for eBGP peering links.

The goal of this procedure is to let the paths being shutdown visible, but with a lower local preference, while alternate paths spread through the iBGP topology. Instead of withdrawing the path, routers of an AS will keep on using it until they become aware of alternate paths.

4.2.1. eBGP g-shut

4.2.1.1. Pre-configuration

On each ASBR supporting the g-shut procedure, an outbound BGP route policy is applied on all iBGP sessions of the ASBR, that:

- matches the g-shut community
- sets the local-pref of the paths tagged with the g-shut community to a low value
- removes the g-shut community from the paths.
- optionally, adds an AS specific g-shut community on these paths to indicate that these are to be withdrawn soon. If some ingress ASBRs reset the local preference attribute, this AS specific g-shut community will be used to override other local preference changes.

Note that in the case where an AS is aggregating multiple routes under a covering prefix, it is recommended to filter out the g-shut community from the resulting aggregate BGP route. By doing so, the setting of the g-shut community on one of the aggregated routes will not let the entire aggregate inherit the community. Not doing so would let the entire aggregate undergo the g-shut behavior.

4.2.1.2. Operations at maintenance time

On the g-shut initiator, upon maintenance time, it is required to:

- apply an outbound BGP route policy on the maintained eBGP session to tag the paths propagated over the session with the g-shut community. This will trigger the BGP implementation to re-advertise all active routes previously advertised, and tag them with the g-shut community.
- apply an inbound BGP route policy on the maintained eBGP session to tag the paths received over the session with the g-shut community.

- o wait for convergence to happen.
- o perform a BGP session shutdown.

4.2.1.3. BGP implementation support for G-Shut

A BGP router implementation MAY provide features aimed at automating the application of the graceful shutdown procedures described above.

Upon a session shutdown specified as graceful by the operator, a BGP implementation supporting a g-shut feature SHOULD:

1. Update all the paths propagated over the corresponding eBGP session, tagging the GSHUT community to them. Any subsequent update sent to the session being gracefully shut down would be tagged with the GSHUT community.
2. Lower the local preference value of the paths received over the eBGP session being shut down, upon their propagation over iBGP sessions. Optionally, also tag these paths with an AS specific g-shut community. Note that alternatively, the local preference of the paths received over the eBGP session can be lowered on the g-shut initiator itself, instead of only when propagating over its iBGP sessions.
3. Optionally shut down the session after a configured time.
4. Prevent the GSHUT community from being inherited by a path that would aggregate some paths tagged with the GSHUT community. This behavior avoids the GSHUT procedure to be applied to the aggregate upon the graceful shutdown of one of its covered prefixes.

A BGP implementation supporting a g-shut feature SHOULD also automatically install the BGP policies that are supposed to be configured, as decribed in Section 4.2.1.1 for sessions over which g-shut is to be supported.

4.2.2. iBGP g-shut

If the iBGP topology is viable after the maintenance of the session, i.e, if all BGP speakers of the AS have an iBGP signaling path for all prefixes advertised on this g-shut iBGP session, then the shutdown of an iBGP session does not lead to transient unreachability.

4.2.3. Router g-shut

In the case of a shutdown of a router, a reconfiguration of the outbound BGP route policies of the g-shut initiator MAY be performed to set a low local-pref value for the paths originated by the g-shut initiator (e.g, BGP aggregates redistributed from other protocols,

including static routes).

This behavior is equivalent to the recommended behavior for paths "redistributed" from eBGP sessions to iBGP sessions in the case of the shutdown of an ASBR.

5. Forwarding modes and transient forwarding loops during convergence

The g-shut procedure or the solutions improving the availability of alternate paths, do not change the fact that BGP convergence and the subsequent FIB updates are runned independently on each router of the ASes. If the AS applying the solution does not rely on encapsulation to forward packets from the Ingress Border Router to the Egress Border Router, then transient forwarding loops and consequent packet losses can occur during the convergence process. If zero LoC is required, encapsulation is required between ASBRs of the AS.

6. Link Up cases

We identify two potential causes for transient packet losses upon an eBGP link up event. The first one is local to the g-no-shut initiator, the second one is due to the BGP convergence following the injection of new best paths within the iBGP topology.

6.1. Unreachability local to the ASBR

An ASBR that selects as best a path received over a newly brought up eBGP session may transiently drop traffic. This can typically happen when the nexthop attribute differs from the IP address of the eBGP peer, and the receiving ASBR has not yet resolved the MAC address associated with the IP address of that "third party" nexthop.

A BGP speaker implementation could avoid such losses by ensuring that "third party" nexthops are resolved before installing paths using these in the RIB.

If the link up event corresponds to an eBGP session that is being manually brought up, over an already up multi-access link, then the operator can ping third party nexthops that are expected to be used before actually bringing the session up, or ping directed broadcast the subnet IP address of the link. By proceeding like this, the MAC addresses associated with these third party nexthops will be resolved by the g-no-shut initiator.

6.2. iBGP convergence

Corner cases leading to LoC can occur during an eBGP link up event.

A typical example for such transient unreachability for a given prefix is the following:

Let’s consider 3 route reflectors RR1, RR2, RR3. There is a full mesh of iBGP session between them.

1. RR1 is initially advertising the current best path to the members of its iBGP RR full-mesh. It propagated that path within its RR full-mesh. RR2 knows only that path towards the prefix.
2. RR3 receives a new best path orginated by the "g-no-shut" initiator, being one of its RR clients. RR3 selects it as best, and propagates an UPDATE within its RR full-mesh, i.e., to RR1 and RR2.
3. RR1 receives that path, reruns its decision process, and picks this new path as best. As a result, RR1 withdraws its previously announced best-path on the iBGP sessions of its RR full-mesh.
4. If, for any reason, RR3 processes the withdraw generated in step 3, before processing the update generated in step 2, RR3 transiently suffers from unreachability for the affected prefix.

The use of [BestExternal] among the RR of the iBGP full-mesh can solve these corner cases by ensuring that within an AS, the advertisement of a new route is not translated into the withdraw of a former route.

Indeed, "best-external" ensures that an ASBR does not withdraw a previously advertised (eBGP) path when it receives an additional, preferred path over an iBGP session. Also, "best-intra-cluster" ensures that a RR does not withdraw a previously advertised (iBGP) path to its non clients (e.g. other RRs in a mesh of RR) when it receives a new, preferred path over an iBGP session.

7. IANA assigned g-shut BGP community

Applying the g-shut procedure is rendered much easier with the use of a single g-shut community value which could be used on all eBGP sessions, for both inbound and outbound signaling. The community value 0xFFFF0000 has been assigned by IANA for this purpose.

For Internet routes, a non transitive extended community will be reserved from the pool defined in [EXT_POOL]. Using such a community

type allows for not leaking graceful signaling out of the AS boundaries, without the need to explicitly configure filters to strip the community off upon path propagation.

8. Security Considerations

By providing the g-shut service to a neighboring AS, an ISP provides means to this neighbor to lower the local-pref value assigned to the paths received from this neighbor.

The neighbor could abuse the technique and do inbound traffic engineering by declaring some prefixes as undergoing a maintenance so as to switch traffic to another peering link.

If this behavior is not tolerated by the ISP, it SHOULD monitor the use of the g-shut community by this neighbor.

ASes using the regular (transitive) g-shut community SHOULD remove the community from neighboring ASes that do not support the g-shut procedure. Doing so prevents malignant remote ASes from using the community through intermediate ASes that do not support the feature, in order to perform inbound traffic engineering. ASes using the non-transitive extended community do not need to do this as the community is non transitive and hence cannot be used by remote ASes.

9. Acknowledgments

The authors wish to thank Olivier Bonaventure and Pradosh Mohapatra for their useful comments on this work.

10. References

[AddPath] D. Walton, A. Retana, and E. Chen, "Advertisement of Multiple Paths in BGP", draft-walton-bgp-add-paths-06.txt (work in progress).

[BestExternal] Marques, P., Fernando, R., Chen, E., and P. Mohapatra, "Advertisement of the best-external route to IBGP", draft-ietf-idr-best-external-00.txt, May 2009.

[REQS] Decraene, B., Francois, P., Pelsser, C., Ahmad, Z., Armengol, A., and T. Takeda, "Requirements for the graceful shutdown of BGP sessions", draft-ietf-grow-bgp-graceful-shutdown-requirements-

06.txt, October 2010.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[EXT_POOL] Decraene, B. and P. Francois, "Assigned BGP extended communities", draft-ietf-idr-reserved-extended-communities-01, May 2011.

[BGPWKC] "http://www.iana.org/assignments/bgp-well-known-communities".

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

Appendix A. Alternative techniques with limited applicability

A few alternative techniques have been considered to provide g-shut capabilities but have been rejected due to their limited applicability. This section describe them for possible reference.

A.1. Multi Exit Discriminator tweaking

The MED attribute of the paths to be avoided can be increased so as to force the routers in the neighboring AS to select other paths.

The solution only works if the alternate paths are as good as the initial ones with respect to the Local-Pref value and the AS Path Length value. In the other cases, increasing the MED value will not have an impact on the decision process of the routers in the neighboring AS.

A.2. IGP distance Poisoning

The distance to the BGP nexthop corresponding to the maintained session can be increased in the IGP so that the old paths will be less preferred during the application of the IGP distance tie-break rule. However, this solution only works for the paths whose alternates are as good as the old paths with respect to their Local-Pref value, their AS Path length, and their MED value.

Also, this poisoning cannot be applied when nexthop self is used as there is no nexthop specific to the maintained session to poison in the IGP.

Authors' Addresses

Pierre Francois
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganese 28918
ES

Email: pierre.francois@imdea.org

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
92794 Issi Moulineaux cedex 9
FR

Email: bruno.decraene@orange.com

Cristel Pelsser
Internet Initiative Japan
Jinbocho Mitsui Bldg.
1-105 Kanda Jinbo-cho
Tokyo 101-0051
JP

Email: pelsser.cristel@iij.ad.jp

Keyur Patel
Cisco Systems
170 West Tasman Dr
San Jose, CA 95134
US

Email: keyupate@cisco.com

Clarence Filsfils
Cisco Systems
De kleetlaan 6a
Diegem 1831
BE

Email: cfilsfil@cisco.com