Network Working Group   Pierre Francois
Internet-Draft   Universite catholique de Louvain
Intended status: Informational   Bruno Decraene
Expires: September 7, 2009   France Telecom
                                                          Cristel Pelsser
                                                         NTT Corporation
                                                       Clarence Filsfils
                                                           Cisco Systems
                                                           March 6, 2009

Graceful BGP session shutdown
draft-francois-bgp-gshut-01

Status of this Memo

Copyright Notice

Abstract

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

Table of Contents

1. Introduction

Routing changes in BGP can be caused by planned, manual, 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 allow routers to keep using old paths until alternate ones are learned, ensuring 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 or BGP implementations.

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 configuration at maintenance time.

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.

Note that for the link-up case, we will refer to these nodes as g-no-shut initiator, and g-no-shut neighbor.

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

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

Affected path / Nominal / pre-convergence path : a BGP path via the

peering link(s) undergoing the maintenance. This path will no longer exist after the shutdown.

Affected prefix : a prefix initially reached via an affected path.

Affected router : a router having an affected prefix.

Backup / alternate / post-convergence path : a path toward an affected prefix that will be selected as the best path by an affected router, when the link is shut down and the BGP convergence is completed.

Transient alternate path : a path towards an affected prefix that may be transiently selected as best by an affected router during the convergence process but that is not a post-convergence path.

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, routers' FIB 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 in routers performing packet lookups in BGP tables [BestExternal] [AddPath] help in reducing the LoC bound with manual shutdown of eBGP sessions.

One solution 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. This simplified procedure would only tackle potential LoC for the inbound traffic.

Using advertise-best-external [BestExternal] on ASBRs and RRs helps in avoiding lack of alternate paths in route reflectors upon a convergence. Hence it reduces the LoC duration for the outbound traffic of the ISP upon an eBGP Session shutdown by reducing the iBGP path hunting.

Still it does not ensure that BGP routers will always have at least one path towards affected prefixes during the convergence following the event. This property may be verified in future revisions of [BestExternal], notably of its Section 4, hence the current proposal will be updated accordingly.

Increasing diversity with [AddPath] might lead to the respect of this property, depending on the path propagation decision process that add-path compliant routers would use.

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, hence the procedure described in Section 4.2.2 should be applied upon the maintenance, even if not required for the outbound traffic.

4.2. Graceful shutdown procedures for eBGP sessions

This section aims at describing a procedure to be applied to reduce the LoC with readily available BGP features, and without assuming a particular iBGP design in the Initiator and Neighbor ASes.

4.2.1. Outbound traffic

This section discusses a mean to render the affected paths less desirable by the BGP decision process of affected routers, still allowing these to be used during the convergence while alternate

paths are propagated to the affected routers.

A decrease of the local-pref value of the affected paths can be issued in order to render the affected paths less preferable, at the highest possible level of the BGP Decision Process.

This operation can be performed by reconfiguring the out-filters associated with the iBGP sessions established by the g-shut initiator.

The modification of the filters MUST supplant any other rule affecting the local-pref value of the old paths.

Compared to using an in-filter of the eBGP session to be shut down, the modification of the out-filters will not let the g-shut initiator switch to another path, as the input to the BGP decision process of that router does not change. As a consequence, the g-shut initiator will not send a withdraw message over its iBGP sessions when it receives an alternate path over an iBGP session. It will however modify the local-pref of the affected paths so that upstream routers will switch to alternate ones.

When the actual shutdown of the session is performed, the g-shut initiator will itself switch to the alternate paths.

4.2.2. Inbound traffic

The solution described for the outbound traffic can be applied at the neighbor AS. This can be done either "manually" or by using a community value dedicated to this task.

4.2.2.1. Phone call

The operator performing the maintenance of the eBGP session can contact the operator at the other side of the peering link, and let him apply the procedure described above for its own outbound traffic.

4.2.2.2. Community tagging

A community value (referred to as GSHUT community in this document) can be agreed upon by neighboring ASes. A path tagged with this community must be considered as soon to be affected by a maintenance operation.

4.2.2.2.1. Pre-Configuration

A g-shut neighbor is pre-configured to set a low local-pref value for the paths received over eBGP sessions which are tagged with the GSHUT

community.

This rule must supplant any other rule affecting the local-pref value of the paths.

This local-pref reconfiguration SHOULD be performed at the out-filters of the iBGP sessions of the g-shut neighbor. That is, the g-shut neighbor does not take into account this low local-pref in its own BGP best path selection. As described in Section 4.2.1 this avoids sending the withdraw messages that can lead to LoC.

4.2.2.2.2. Operational action upon maintenance

Upon the manual shutdown, the output filter associated with the maintained eBGP session will be modified on the g-shut initiator so as to tag all the paths advertised over the session with the GSHUT community.

4.2.2.2.3. Transitivity of the community

If the GSHUT community is an extended community, it SHOULD be set non transitive.

If a normal community is used, this community SHOULD be removed from the path by the ASBR of the peer receiving it. If not, the GSHUT community MAY be removed from the path by all the ASBRs of the neighboring AS, before propagating the path to other peers.

Not propagating the community further in the Internet reduces the amount of BGP churn and avoids rerouting in distant ASes that would also recognize this community value. In other words, it helps concealing the convergence at the maintenance location.

There are cases where an interdomain exploration is to be performed to recover the reachability, e.g., in the case of a shutdown in confederations where the alternate paths will be found in another AS of the confederation. In such scenarios, the community value SHOULD be allowed to transit through the confederation but MAY be removed from the paths advertised outside of the confederation.

When the local-pref value of a path is conserved upon its propagation from one AS of the confederation to the other, there is no need to have the GSHUT community be propagated throughout that confederation.

4.2.2.2.4. Easing the configuration for G-SHUT

From a configuration burden viewpoint, it would be much easier to reserve a value for the GSHUT community.

First, on the g-shut initiator, an operator would have a single configuration rule to be applied at the maintenance time, which would not depend on the identity of its peer. This would make the maintenance operations less error prone.

Second, on the g-shut neighbor, a simple filter related to g-shut can be applied to all iBGP sessions. Additionnaly, this filter doesn't need to be updated each time neighboring ASes are added or removed.

4.3. Graceful shutdown procedures for iBGP sessions

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.

However, in the case of a shutdown of a router, a reconfiguration of the out-filters of the g-shut initiator should 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 forwarding loops

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, even if the procedure described above is applied. Hence if zero LoC is required, encapsulation is required between ASBRs of the AS.

Using the out-filter reconfiguration avoids the forwarding loops between the g-shut initiator and its directly connected upstream neighbors. Indeed, when this reconfiguration is applied, the g-shut initiator keeps using its own external path and lets the upstream routers converge to the alternate ones. During this phase, no forwarding loops can occur between the g-shut initiator and its upstream neighbors as the g-shut initiator keeps using the affected paths via its eBGP peering links. When all the upstream routers have switched to alternate paths, the transition performed by the g-shut initiator when the session is actually shut down, will be loopfree. Transient forwarding loops between other routers will not be avoided

with this procedure.

6. Dealing with Internet policies

A side gain of the maintenance solution is that it can also reduce the churn implied by a shutdown of an eBGP session.

For this, it is recommended to apply the filters modifying the local-pref value of the paths to values strictly lower but as close as possible to the local-pref values of the post-convergence paths.

For example, if an eBGP link is shut down between a provider and one of its customers, and another link with this customer remains active, then the value of the local-pref of the old paths SHOULD be decreased to the smallest possible value of the 'customer' local_pref range, minus 1. Thus, routers will not transiently switch to paths received from shared-cost peers or providers, which could lead to the propagation of withdraw messages over eBGP sessions with shared-cost peers and providers.

Proceeding like this reduces both BGP churn and traffic shifting as routers will less likely switch to transient paths.

In the above example, it also prevents transient unreachabilities in the neighboring AS that are due to the sending of "abrupt" withdraw messages to shared-cost peers and providers.

7. Effect of the g-shut procedure on the convergence

This section describes the effect of applying the solution.

7.1. Maintenance of an eBGP session

This section describes the effect of applying the solution for the shutdown of an eBGP session.

7.1.1. Propagation on the other eBGP sessions of the g-shut initiator

Nothing is propagated on the other eBGP sessions when the out-filters reconfiguration step is applied. The reconfiguration is indeed only defined for its iBGP sessions.

The reconfiguration of the iBGP out-filters will trigger the reception of alternate paths at the g-shut initiator. As the eBGP in-filters have not been modified at that step, the old paths are still preferred by the g-shut initiator.

7.1.2. Propagation on the other iBGP sessions of the g-shut initiator

During the out-filter reconfiguration, path updates are propagated with a reduced local-pref value for the affected paths. As a consequence, Route Reflectors and distant ASBRs select and propagate alternate paths through the iBGP topology as they no longer select the old paths as best.

When the shut-down is performed, for each affected prefix, the g-shut initiator propagates on its iBGP sessions:

. The alternate path, if the best path was received over an eBGP sessions.

. A withdraw, if the best path was received over an iBGP sessions.

7.1.3. Propagation of updates in an iBGP full-mesh

No transient LoC can occur if a reconfiguration of the iBGP out-filters on the g-shut initiator is performed.

7.1.4. Propagation of updates from iBGP to iBGP in a RR hierarchy

Upon the reception of the update of a primary path with a lower local-pref value from a client, a Route Reflector RR1 will either propagate the update, or select an alternate path, depending on the fact that the updated primary path is still the best one w.r.t. the state of the Adj-Rib-In of RR1.

If the updated primary path is still the best, then the RR will propagate an update for this path to the iBGP neighbors to which it previously advertised the path. Hence it cannot cause transient lack of path in the Adj-Rib-In of its iBGP neighbors.

If an alternate path is picked, and this path was also originated by a client of RR1, an update will also be propagated to the same neighbors as the one to which the primary path was initially propagated. Hence it cannot cause transient lack of path in the Adj-Rib-In of its iBGP neighbors.

If an alternate path is picked, and this path was received from a member of its Route-Reflector iBGP full-mesh, then a withdraw message is sent. As the alternate path has been sent over each session of the iBGP full-mesh, the propagation of a withdraw for the primary path of RR1 is done to routers that are expected to know the alternate path picked by RR1.

The following example describes a situation where some corner case

timings could lead to transient unreachability from some members of the iBGP full-mesh.

1. A Route Reflector RR1 only knew about the primary path upon the shutdown.

2. A member of its RR full-mesh, RR2, propagates an update of the old path with a lower local-pref.

3. Another member of its RR full-mesh, RR3 processes the update, selects an alternate path, and propagates an update in the mesh.

4. RR2 receives the alternate path, selects it as best, and hence withdraws the updated old path on the iBGP sessions of the mesh.

5. If for any reason, RR1 receives and processes the withdraw generated in step 4 before processing the update generated in step 3, RR1 transiently suffers from unreachability for the affected prefix.

In such corner cases, the solution improves the iBGP convergence behavior/LoC but does not ensure 0 packet loss, as we cannot define a simple solution relying only on a reconfiguration of the filters of the g-shut initiator. Improving the availability of alternate paths in Route Reflectors, using [BestExternal], or [AddPath], seems to be the most pragmatic solution to these corner cases.

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

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.

7.2. Maintenance of an iBGP session

If the shutdown does not temper with the viability of the iBGP topology, the described procedure is sufficient to avoid LoC.

7.3. Applicability of the g-shut procedure

The applicability of the procedure described in this draft to the cases presented in [REQS] can be shown by combining the effects described in this section. A complete case by case analysis will be provided in the next versions of the draft.

7.4. Summary of operations

This section summarizes the configurations and actions to be performed to support the g-shut procedure for eBGP peering links.

7.4.1. Pre-configuration

On each ASBR supporting the g-shut procedure, set-up an out-filter applied on all iBGP sessions of the ASBR, that :

. sets the local-pref of the paths tagged with the g-shut community to a low value

. removes the g-shut community from the path.

7.4.2. Operations at maintenance time

On the g-shut initiator :

. Apply an in-filter on the maintained BGP session to tag the paths received over the session with the g-shut community.

. Apply an out-filter on the maintained BGP session to tag the paths propagated over the session with the g-shut community.

. Wait for convergence to happen.

. Perform a BGP session shutdown.

8. 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-shut initiator, the second one is due to the BGP convergence following the injection of new best paths within the iBGP topology.

8.1. 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.

8.2. iBGP convergence

Similar corner cases as described in Section 7.1.4 for the link down case, can occur during an eBGP link up event.

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

1. A Route Reflector, 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. Another route reflector of the full-mesh, RR2, knows only that path towards the prefix.
2. A third Route Reflector of the RR full-mesh, 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.

9. 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.

9.1. In-filter reconfiguration

An In-filter reconfiguration on the eBGP session undergoing the maintenance could be performed instead of out-filter reconfigurations on the iBGP sessions of the g-shut initiator.

Upon the application of the maintenance procedure, if the g-shut initiator has an alternate path in its Adj-Rib-In, it will switch to it directly.

If this new path was advertised by an eBGP neighbor of the g-shut initiator, the g-shut initiator will send a BGP Path Update message advertising the new path over its iBGP and eBGP sessions.

If this new path was received over an iBGP session, the g-shut initiator will select that path and withdraw the previously advertised path over its non-client iBGP sessions. There can be iBGP topologies where the iBGP peers of the g-shut initiator do not know an alternate path, and hence may drop traffic.

Also, applying an In-filter reconfiguration on the eBGP session undergoing the maintenance may lead to transient LoC, in full-mesh iBGP topologies if

a. An ASBR of the initiator AS, ASBR1 did not initially select its own external path as best, and

b. An ASBR of the initiator AS, ASBR2 advertises a new path along its iBGP sessions upon the reception of ASBR1's update following the in-filter reconfiguration on the g-shut initiator, and

c. ASBR1 receives the update message, runs its Decision Process and hence withdraws its external path after having selected ASBR2's path as best, and

d. An impacted router of the AS processes the withdraw of ASBR1 before processing the update from ASBR2.

Applying a reconfiguration of the out-filters prevents such transient unreachabilities.

Indeed, when the g-shut initiator propagates an update of the old path first, the withdraw from ASBR2 does not trigger unreachability in other nodes, as the old path is still available. Indeed, even though it receives alternate paths, the g-shut initiator keeps using its old path as best as the in-filter of the maintained eBGP session has not been modified yet.

Applying the out-filter reconfiguration also prevents packet loops between the g-shut initiator and its direct neighbors when encapsulation is not used between the ASBRs of the AS.

9.2. 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.

9.3. 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.

10. IANA considerations

Applying the g-shut procedure is rendered much easier with a reserved g-shut community value. Hence this draft suggests to reserve a community value, e.g., 0xFFFF0000, for this purpose.

11. 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.

12. Acknowledgments

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

13. 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-marques-idr-best-external-00.txt, July 2008.

[REQS]   Decraene, B., Francois, P., Pelsser, C., and Z. Ahmad, "Requirements for the graceful shutdown of BGP sessions", draft-decraene-bgp-graceful-shutdown-requirements-00.txt , December 2007.

Authors' Addresses

Pierre Francois
Universite catholique de Louvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348
BE

Email: pierre.francois@uclouvain.be
URI: http://inl.info.ucl.ac.be/pfr

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

Email: bruno.decraene@orange-ftgroup.com

Cristel Pelsser
NTT Corporation
9-11, Midori-Cho 3 Chrome
Musashino-Shi, Tokyo 180-8585
JP

Email: pelsser.cristel@lab.ntt.co.jp

Clarence Filsfils
Cisco Systems
De kleetlaan 6a
Diegem 1831
BE

Email: cfilsfil@cisco.com