Network Working Group   Pierre Francois
Internet-Draft   Individual Contributor
Intended status: Informational   B. Decraene
Expires: December 27, 2017   Orange
                                                           C. Pelsser
                                                 Strasbourg University
                                                             K. Patel
                                                          Arrcus, Inc.
                                                          C. Filsfils
                                                        Cisco Systems
                                                        June 25, 2017

                    Graceful BGP session shutdown
                     draft-ietf-grow-bgp-gshut-08

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. It defines a
   well-known BGP community, called g-shut, to signal the graceful
   shutdown of paths to other Autonomous Systems.

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 December 27, 2017.

Copyright Notice

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
[RFC6198] at best, without changing the BGP protocol.

This document defines a well-known community [RFC1997], called
g-shut, for the purpose of reducing the management overhead of
gracefully shutting down BGP sessions. The well-known community
allows implementers to provide an automated 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 has a BGP session, to be shutdown,
with the g-shut initiator.

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
[I-D.ietf-idr-best-external] and [RFC7911] 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 for
the graceful shutdown of BGP sessions.

The goal of this procedure is to let, in both ASes, the paths being
shutdown visible, but with a lower LOCAL_PREF value, 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

This section describes configurations and actions to be performed for
the graceful shutdown of EBGP peering links.

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:

o matches the g-shut community

o sets the LOCAL_PREF attribute of the paths tagged with the g-shut
   community to a low value

o removes the g-shut community from the paths.

o 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_PREF attribute, this AS specific g-shut
   community will be used to override other LOCAL_PREF 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:

o apply an outbound BGP route policy on the EBGP session to be
   shutdown. This policy tags 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.

o 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 shutdown the EBGP session, optionally using
   [I-D.ietf-idr-shutdown] to communicate the reason of the 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. On the EBGP side, update all the paths propagated over the
       corresponding EBGP session, tagging the g-shut community to them.
       Any subsequent update sent over the session being gracefully shut
       down would be tagged with the g-shut community.

   2. On the iBGP side, lower the LOCAL_PREF 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.

   3. Optionally shut down the session after a configured time.

   4. Prevent the g-shut 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 described in Section 4.2.1.1 for sessions over which
   g-shut is to be supported.

4.2.2. **iBGP g-shut**

   For the shutdown of an iBGP session, provided 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. As a consequence, no specific
   g-shut action is required.

4.2.3. **Router g-shut**

   In the case of a shutdown of the whole router, in addition to the
   g-shut of all EBGP sessions, there is a need to g-shut the routes
   originated by this router (e.g, BGP aggregates redistributed from
   other protocols, including static routes). This can be performed by
   tagging such routes with the g-shut community.

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 run 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 toward the prefix.

2. RR3 receives a new best path originated 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 [I-D.ietf-idr-best-external] 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 Considerations

The IANA has assigned the community value 0xFFFF0000 to the planned-
shut community in the "BGP Well-known Communities" registry. IANA is
requested to change the name planned-shut to g-shut and set this
document as the reference.

8. Security Considerations

By providing the g-shut service to a neighboring AS, an ISP provides
means to this neighbor and possibly its downstream ASes 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.

[9](#). Acknowledgments

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

[10](#). References

[10.1](#). Normative References

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", [RFC 1997](#), DOI 10.17487/RFC1997, August 1996, <[http://www.rfc-editor.org/info/rfc1997](#)>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <[http://www.rfc-editor.org/info/rfc2119](#)>.

[RFC6198] Decraene, B., Francois, P., Pelsser, C., Ahmad, Z., Elizondo Armengol, A., and T. Takeda, "Requirements for the Graceful Shutdown of BGP Sessions", [RFC 6198](#), DOI 10.17487/RFC6198, April 2011, <[http://www.rfc-editor.org/info/rfc6198](#)>.

[10.2](#). Informative References

[I-D.ietf-idr-best-external]
           Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", [draft-ietf-idr-best-external-05](#) (work in progress), January 2012.

[I-D.ietf-idr-shutdown]
           Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", [draft-ietf-idr-shutdown-10](#) (work in progress), June 2017.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", [RFC 7911](#), DOI 10.17487/RFC7911, July 2016, <[http://www.rfc-editor.org/info/rfc7911](#)>.

[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
   Individual Contributor

   Email: pfrpfr@gmail.com

   Bruno Decraene
   Orange

   Email: bruno.decraene@orange.com

   Cristel Pelsser
   Strasbourg University

   Email: pelsser@unistra.fr

   Keyur Patel
   Arrcus, Inc.

   Email: keyur@arrcus.com

Clarence Filsfils
Cisco Systems

Email: cfilsfil@cisco.com