Internet Engineering Task Force (IETF)
Request for Comments: 8326
Category: Standards Track
ISSN: 2070-1721

P. Francois, Ed.
Individual Contributor
B. Decraene, Ed.
Orange
C. Pelsser
Strasbourg University
K. Patel
Arrcus, Inc.
C. Filsfils
Cisco Systems

March 2018

Graceful BGP Session Shutdown

Abstract

This document standardizes a new well-known BGP community, GRACEFUL_SHUTDOWN, to signal the graceful shutdown of paths. This document also describes operational procedures that use this well-known community to reduce the amount of traffic lost when BGP peering sessions are about to be shut down deliberately, e.g., for planned maintenance.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc8326.

Table of Contents

1. Introduction

Routing changes in BGP can be caused by planned maintenance operations. This document defines a well-known community [RFC1997], called GRACEFUL_SHUTDOWN, 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.

This document discusses operational procedures to be applied in order to reduce or eliminate loss of packets during a maintenance operation. Loss comes from transient lack of reachability during BGP convergence that follows the shutdown of an EBGP peering session between two Autonomous System Border Routers (ASBRs).

This document presents procedures for the cases where the forwarding plane is impacted by the maintenance, hence for 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. In both Autonomous Systems (ASes), these procedures trigger rerouting to alternate paths if they exist within the AS while allowing the use of the old path until alternate ones are learned. 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] as best possible without changing BGP.

Other maintenance cases, such as the shutdown of an IBGP session or the establishment of an EBGP session, are out of scope for this document. For informational purposes, they are briefly discussed in Appendix C.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

graceful shutdown initiator
: A router on which the session shutdown is performed for the maintenance.

graceful shutdown receiver
: A router that has a BGP session to be shut down with the graceful shutdown initiator.

3. Packet Loss upon Manual EBGP Session Shutdown

Packets can be lost during the BGP convergence following a manual shut down of an EBGP session for two reasons.

First, some routers can have no path toward 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 extension defined in [RFC7911] is not used and a) the paths are not selected as best by the ASBRs that receive them on an EBGP session or b) Route Reflectors do not propagate the paths further in the IBGP topology because they do not select them as best.

Second, the FIB can be inconsistent between routers within the AS, and packets toward affected prefixes can loop and be dropped unless encapsulation is used within the AS.

This document only addresses the first reason.

4. Procedure for EBGP Graceful Shutdown

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

The goal of this procedure is to retain the paths to be shut down between the peers, but with a lower LOCAL_PREF value, allowing the paths to remain in use while alternate paths are selected and propagated, rather than simply withdrawing the paths. The LOCAL_PREF value SHOULD be lower than any of the alternative paths. The RECOMMENDED value is 0.

Note that some alternative techniques with limited applicability are discussed in Appendix A for informational purposes.

4.1. Pre-configuration

On each ASBR supporting the graceful shutdown receiver procedure, an inbound BGP route policy is applied on all EBGP sessions of the ASBR. That policy:

- matches the GRACEFUL_SHUTDOWN community.
- sets the LOCAL_PREF attribute of the paths tagged with the GRACEFUL_SHUTDOWN community to a low value.

For informational purposes, examples of configurations are provided in Appendix B.

4.2. Operations at Maintenance Time

On the graceful shutdown initiator, at maintenance time, the operator:

- applies an outbound BGP route policy on the EBGP session to be shutdown. This policy tags the paths propagated over the session with the GRACEFUL_SHUTDOWN community. This will trigger the BGP implementation to re-advertise all active routes previously advertised and tag them with the GRACEFUL_SHUTDOWN community.
- applies an inbound BGP route policy on the EBGP session to be shutdown. This policy tags the paths received over the session with the GRACEFUL_SHUTDOWN community and sets LOCAL_PREF to a low value.
- waits for route re-advertisement over the EBGP session and for BGP routing convergence on both ASBRs.
- shuts down the EBGP session, optionally using [RFC8203] to communicate the reason for the shutdown.

In the case of a shutdown of the whole router, in addition to the graceful shutdown of all EBGP sessions, there is a need to gracefully shut down the routes originated by this router (e.g., BGP aggregates redistributed from other protocols, including static routes). This can be performed by tagging these routes with the GRACEFUL_SHUTDOWN community and setting LOCAL_PREF to a low value.

4.3. BGP Implementation Support for Graceful Shutdown

BGP Implementers SHOULD provide configuration knobs that utilize the GRACEFUL_SHUTDOWN community to inform BGP neighbors in preparation for an impending neighbor shutdown. Implementation details are outside the scope of this document.

5. IANA Considerations

IANA previously assigned the community value 0xFFFF0000 to the 'planned-shut' community in the "BGP Well-known Communities" registry. IANA has changed the name 'planned-shut' to 'GRACEFUL_SHUTDOWN' and updated the reference to point to this document.

6. Security Considerations

By providing the graceful shutdown 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 that some prefixes are undergoing 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 graceful shutdown community.

7. References

7.1. Normative References

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://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, <https://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, <https://www.rfc-editor.org/info/rfc6198>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[BEST-EXTERNAL] Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", Work in Progress, draft-ietf-idr-best-external-05, January 2012.

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

[RFC8203] Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", RFC 8203, DOI 10.17487/RFC8203, July 2017, <https://www.rfc-editor.org/info/rfc8203>.

Appendix A. Alternative Techniques with Limited Applicability

A few alternative techniques have been considered to provide graceful shutdown capabilities but have been rejected due to their limited applicability. This section describes these techniques for possible reference.

A.1. Multi-Exit Discriminator Tweaking

The Multi-Exit Discriminator (MED) attribute of the paths to be avoided can be increased to influence 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 NEXT_HOP 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 BGP "NEXT_HOP self" is used, as there is no BGP NEXT_HOP specific to the maintained session to poison in the IGP.

Appendix B. Configuration Examples

This appendix is non-normative.

This appendix includes examples of routing policy configurations to honor the GRACEFUL_SHUTDOWN well-known BGP community.

B.1. Cisco IOS XR

```
community-set comm-graceful-shutdown
  65535:0
end-set
!
route-policy AS64497-ebgp-inbound
  ! normally this policy would contain much more
  if community matches-any comm-graceful-shutdown then
    set local-preference 0
  endif
end-policy
!
router bgp 64496
 neighbor 2001:db8:1:2::1
  remote-as 64497
  address-family ipv6 unicast
   send-community-ebgp
   route-policy AS64497-ebgp-inbound in

  !
 !
!
```

B.2. BIRD

```
function honor_graceful_shutdown() {
    if (65535, 0) ~ bgp_community then {
        bgp_local_pref = 0;
    }
}
filter AS64497_ebgp_inbound
{
        # normally this policy would contain much more
        honor_graceful_shutdown();
}
protocol bgp peer_64497_1 {
    neighbor 2001:db8:1:2::1 as 64497;
    local as 64496;
    import keep filtered;
    import filter AS64497_ebgp_inbound;
}
```

B.3. OpenBGPD

```
AS 64496
router-id 192.0.2.1
neighbor 2001:db8:1:2::1 {
        remote-as 64497
}
# normally this policy would contain much more
match from any community GRACEFUL_SHUTDOWN set { localpref 0 }
```

Appendix C. Beyond EBGP Graceful Shutdown

C.1. IBGP Graceful Shutdown

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 graceful shutdown IBGP session), then the shutdown of an IBGP session does not lead to transient unreachability. As a consequence, no specific graceful shutdown action is required.

C.2. EBGP Session Establishment

We identify two potential causes for transient packet losses upon the establishment of an EBGP session. The first one is local to the startup initiator; the second one is due to the BGP convergence following the injection of new best paths within the IBGP topology.

C.2.1. Unreachability Local to the ASBR

An ASBR that selects a path received over a newly established EBGP session as the best path may transiently drop traffic. This can typically happen when the NEXT_HOP 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 NEXT_HOP.

A BGP speaker implementation MAY avoid such losses by ensuring that third-party NEXT_HOPs are resolved before installing paths using these NEXT_HOPs in the RIB.

Alternatively, the operator (script) MAY ping third-party NEXT_HOPs that are expected to be used prior to establishing the session. By proceeding like this, the MAC addresses associated with these third-party NEXT_HOPs are resolved by the startup initiator.

C.2.2. IBGP Convergence

During the establishment of an EBGP session, in some corner cases, a router may have no path toward an affected prefix, leading to loss of connectivity.

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

Consider three Route Reflectors (RR): RR1, RR2, RR3. There is a full mesh of IBGP sessions 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 startup initiator, which is 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 [RFC7911] or [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, advertising the best external route ensures that an ASBR does not withdraw a previously advertised (EBGP) path when it receives an additional, preferred path over an IBGP session. Also, advertising the best intra-cluster route ensures that an 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.

Acknowledgments

The authors wish to thank Olivier Bonaventure, Pradosh Mohapatra, Job Snijders, John Heasley, and Christopher Morrow for their useful comments.

Authors' Addresses

Pierre Francois (editor)
Individual Contributor

Email: pfrpfr@gmail.com

Bruno Decraene (editor)
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