Internet Engineering Task Force (IETF)
Request for Comments: 8327
BCP: 214
Category: Best Current Practice
ISSN: 2070-1721

W. Hargrave
LONAP
M. Griswold
20C
J. Snijders
NTT
N. Hilliard
INEX

March 2018

Mitigating the Negative Impact of Maintenance through BGP Session Culling

Abstract

This document outlines an approach to mitigate the negative impact on networks resulting from maintenance activities. It includes guidance for both IP networks and Internet Exchange Points (IXPs). The approach is to ensure BGP-4 sessions that will be affected by maintenance are forcefully torn down before the actual maintenance activities commence.

Status of This Memo

This memo documents an Internet Best Current Practice.

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 BCPs 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/rfc8327.

Table of Contents

1. Introduction

BGP Session Culling is the practice of ensuring BGP sessions are forcefully torn down before maintenance activities on a lower-layer network commence -- activities that otherwise would affect the flow of data between the BGP speakers. BGP Session Culling is the practice of ensuring BGP sessions are forcefully torn down before commencing maintenance activities (that otherwise would affect the flow of data between the BGP speakers) on a lower-layer network.

BGP Session Culling minimizes the amount of disruption that lower-layer network maintenance activities cause, by making BGP speakers preemptively converge onto alternative paths while the lower-layer network's forwarding plane remains fully operational.

The grace period required for a successful application of BGP Session Culling is the sum of the time needed to detect the loss of the BGP session plus the time required for the BGP speaker to converge onto alternative paths. The first value is often governed by the BGP Hold Timer (see Section 6.5 of [RFC4271]), which is commonly between 90 and 180 seconds. The second value is implementation specific, but it could be as much as 15 minutes when a router with a slow control plane is receiving a full set of Internet routes.

Throughout this document, the "Caretaker" is defined to be in control of the lower-layer network, while "Operators" directly administrate the BGP speakers. Operators and Caretakers implementing BGP Session Culling are encouraged to avoid using a fixed grace period, and instead to monitor forwarding-plane activity while the culling is taking place and to consider it complete once traffic levels have dropped to a minimum (Section 3.3).

2. Requirements Language

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.

3. BGP Session Culling

From the viewpoint of the Operator, there are two types of BGP Session Culling:

Voluntary BGP Session Teardown: The Operator initiates the teardown of the potentially affected BGP session by issuing an Administrative Shutdown.

Involuntary BGP Session Teardown: The Caretaker of the lower-layer network disrupts (higher-layer) BGP control-plane traffic, causing the BGP Hold Timers of the affected BGP session to expire, subsequently triggering rerouting of end-user traffic.

3.1. Voluntary BGP Session Teardown Recommendations

Before an Operator commences activities that can cause disruption to the flow of data through the lower-layer network, an Operator can reduce loss of traffic by issuing an administrative shutdown to all BGP sessions running across the lower-layer network and wait a few minutes for data-plane traffic to subside.

While architectures exist to facilitate quick network reconvergence (such as BGP Prefix Independent Convergence (PIC) [BGP_PIC]), an Operator cannot assume the remote side has such capabilities. As such, a grace period between the Administrative Shutdown and the impacting maintenance activities is warranted.

After the maintenance activities have concluded, the Operator is expected to restore the BGP sessions to their original Administrative state.

3.1.1. Maintenance Considerations

Initiators of the Administrative Shutdown MAY consider using Graceful Shutdown [RFC8326] to facilitate smooth drainage of traffic prior to session tear down, and the Shutdown Communication [RFC8203] to inform the remote side on the nature and duration of the maintenance activities.

3.2. Involuntary BGP Session Teardown Recommendations

In the case where multilateral interconnection between BGP speakers is facilitated through a switched Layer 2 fabric, such as commonly seen at Internet Exchange Points (IXPs), different operational considerations can apply.

Operational experience shows that many Operators are unable to carry out the Voluntary BGP Session Teardown recommendations, because of the operational cost and risk of coordinating the two configuration changes required. This has an adverse affect on Internet performance.

In the absence of notifications from the lower layer (e.g., Ethernet link down) consistent with the planned maintenance activities in a switched Layer 2 fabric, the Caretaker of the fabric could choose to cull BGP sessions on behalf of the Operators connected to the fabric.

Such culling of control-plane traffic will preempt the loss of end-user traffic by causing the expiration of BGP Hold Timers ahead of the moment where the expiration would occur without intervention from the fabric's Caretaker.

In this scenario, BGP Session Culling is accomplished as described in the next subsection, through the application of a combined Layer 3 and Layer 4 (Layer 3/4) packet filter deployed in the Caretaker's switched fabric.

3.2.1. Packet-Filter Considerations

The peering LAN prefixes used by the IXP form the control plane, and the following considerations apply to the packet-filter design:

- The packet filter MUST only affect BGP traffic specific to the Layer 2 fabric, i.e., traffic forming part of the control plane of the system described, rather than multihop BGP traffic that merely transits.
- The packet filter MUST only affect BGP, i.e., TCP port 179.
- The packet filter SHOULD make provision for the bidirectional nature of BGP, i.e., sessions may be established in either direction.
- The packet filter MUST affect all Address Family Identifiers.

Appendix A contains examples of correct packet filters for various platforms.

3.2.2. Hardware Considerations

Not all hardware is capable of deploying combined Layer 3/4 filters on Layer 2 ports; even on platforms that claim support for such a feature, limitations may exist or hardware resource allocation failures may occur during filter deployment, which may cause unexpected results. These problems may include:

- Platform inability to apply Layer 3/4 filters on ports that already have Layer 2 filters applied.
- Layer 3/4 filters supported for IPv4 but not for IPv6.
- Layer 3/4 filters supported on physical ports, but not on IEEE 802.1AX Link Aggregate ports [IEEE802.1AX].

- Failure of the Caretaker to apply filters to all IEEE 802.1AX Link Aggregate ports [IEEE802.1AX].
- Limitations in Access Control List (ACL) hardware mechanisms causing filters not to be applied.
- Fragmentation of ACL lookup memory causing transient ACL application problems that are resolved after ACL removal/reapplication.
- Temporary service loss during hardware programming.
- Reduction in hardware ACL capacity if the platform enables lossless ACL application.

It is advisable for the Caretaker to be aware of the limitations of their hardware and to thoroughly test all complicated configurations in advance to ensure that problems don't occur during production deployments.

3.3. Procedural Considerations

The Caretaker of the lower-layer network can monitor data-plane traffic (e.g., interface counters) and carry out the maintenance without impact to traffic once session culling is complete.

It is recommended that the packet filters be deployed for the duration of the maintenance only and be removed immediately after the maintenance is completed. To prevent unnecessary troubleshooting, it is RECOMMENDED that Caretakers notify the affected Operators before the maintenance takes place and make it explicit that the Involuntary BGP Session Culling methodology will be applied.

4. Security Considerations

There are no security considerations.

5. IANA Considerations

This document has no actions for IANA.

6. References

6.1. Normative References

[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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[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>.

6.2. Informative References

[BGP_PIC] Bashandy, A., Ed., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", Work in Progress, draft-ietf-rtgwg-bgp-pic-06, November 2017.

[IEEE802.1AX] IEEE, "IEEE Standard for Local and metropolitan area networks -- Link Aggregation", IEEE Std 802.1AX-2014, DOI 10.1109/IEEESTD.2014.7055197, December 2014, <http://ieeexplore.ieee.org/servlet/opac?punumber=6997981>.

[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>.

[RFC8326] Francois, P., Ed., Decraene, B., Ed., Pelsser, C., Patel, K., and C. Filsfils, "Graceful BGP Session Shutdown", RFC 8326, DOI 10.17487/8326, March 2018, <https://www.rfc-editor.org/info/rfc8326>.

Appendix A. Example Packet Filters

This section includes examples of packet filters performing Involuntary BGP Session Teardown at an IXP using peering LAN prefixes 192.0.2.0/24 and 2001:db8:2::/64 as its control plane.

A repository of configuration examples for a number of assorted platforms can be found at <https://github.com/bgp/bgp-session-culling-config-examples>.

A.1. Example Configuration for Cisco IOS, IOS XR, and Arista EOS

```
ipv6 access-list acl-ipv6-permit-all-except-bgp
   10 deny tcp 2001:db8:2::/64 eq bgp 2001:db8:2::/64
   20 deny tcp 2001:db8:2::/64 2001:db8:2::/64 eq bgp
   30 permit ipv6 any any
!
ip access-list acl-ipv4-permit-all-except-bgp
   10 deny tcp 192.0.2.0/24 eq bgp 192.0.2.0/24
   20 deny tcp 192.0.2.0/24 192.0.2.0/24 eq bgp
   30 permit ip any any
!
interface Ethernet33
   description IXP Participant Affected by Maintenance
   ip access-group acl-ipv4-permit-all-except-bgp in
   ipv6 access-group acl-ipv6-permit-all-except-bgp in
!
```

A.2. Example Configuration for Nokia SR OS

```
ip-filter 10 create
    filter-name "ACL IPv4 Permit All Except BGP"
    default-action forward
    entry 10 create
        match protocol tcp
            dst-ip 192.0.2.0/24
            src-ip 192.0.2.0/24
            port eq 179
        exit
        action
            drop
        exit
    exit
exit

ipv6-filter 10 create
    filter-name "ACL IPv6 Permit All Except BGP"
    default-action forward
    entry 10 create
        match next-header tcp
            dst-ip 2001:db8:2::/64
            src-ip 2001:db8:2::/64
            port eq 179
        exit
        action
            drop
        exit
    exit
exit

interface "port-1/1/1"
    description "IXP Participant Affected by Maintenance"
    ingress
        filter ip 10
        filter ipv6 10
    exit
exit
```

Acknowledgments

The authors would like to thank the following people for their contributions to this document: Saku Ytti, Greg Hankins, James Bensley, Wolfgang Tremmel, Daniel Roesen, Bruno Decraene, Tore Anderson, John Heasley, Warren Kumari, Stig Venaas, and Brian Carpenter.

Authors' Addresses

Will Hargrave
LONAP Ltd
5 Fleet Place
London EC4M 7RD
United Kingdom

Email: will@lonap.net

Matt Griswold
20C
1658 Milwaukee Ave # 100-4506
Chicago, IL 60647
United States of America

Email: grizz@20c.com

Job Snijders
NTT Communications
Theodorus Majofskistraat 100
Amsterdam 1065 SZ
The Netherlands

Email: job@ntt.net

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
Ireland

Email: nick@inex.ie