

Global Routing Operations
Internet-Draft
Intended status: Best Current Practice
Expires: September 28, 2017

W. Hargrave
LONAP
M. Griswold
20C
J. Snijders
NTT
N. Hilliard
INEX
March 27, 2017

Mitigating Negative Impact of Maintenance through BGP Session Culling draft-iops-grow-bgp-session-culling-01

Abstract

This document outlines an approach to mitigate 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 affected by the maintenance are forcefully torn down before the actual maintenance activities commence.

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 September 28, 2017.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

- [1.](#) Introduction [2](#)
- [2.](#) Requirements Language [3](#)
- [3.](#) BGP Session Culling [3](#)
 - [3.1.](#) Voluntary BGP Session Teardown Recommendations [3](#)
 - [3.1.1.](#) Maintenance Considerations [4](#)
 - [3.2.](#) Involuntary BGP Session Teardown Recommendations [4](#)
 - [3.2.1.](#) Packet Filter Considerations [4](#)
 - [3.2.2.](#) Hardware Considerations [5](#)
 - [3.3.](#) Procedural Considerations [6](#)
- [4.](#) Acknowledgments [6](#)
- [5.](#) Security Considerations [6](#)
- [6.](#) IANA Considerations [6](#)
- [7.](#) References [6](#)
 - [7.1.](#) Normative References [6](#)
 - [7.2.](#) Informative References [6](#)
- [Appendix A.](#) Example packet filters [7](#)
 - [A.1.](#) Cisco IOS, IOS XR & Arista EOS Firewall Example Configuration [7](#)
 - [A.2.](#) Nokia SR OS Filter Example Configuration [7](#)
- Authors' Addresses [8](#)

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, which otherwise would affect the flow of data between the BGP speakers.

BGP Session Culling ensures that lower layer network maintenance activities cause the minimum possible amount of disruption, by causing BGP speakers to preemptively gracefully 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 governed by the BGP Hold Timer ([section 6.5 of \[RFC4271\]](#)), commonly between 90 and 180 seconds, The

second value is implementation specific, but 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 the operator 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, but instead monitor forwarding plane activity while the culling is taking place and 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", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC 2119](#) [[RFC2119](#)].

3. BGP Session Culling

From the viewpoint of the IP network operator, there are two types of BGP Session Culling:

Voluntary BGP Session Teardown: The operator initiates the tear down of the potentially affected BGP session by issuing an Administrative Shutdown.

Involuntary BGP Session Teardown: The caretaker of the lower layer network disrupts BGP control-plane traffic in the upper layer, 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 which can cause disruption to the flow of data through the lower layer network, an operator can reduce loss of traffic by issuing an Administratively 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 PIC [[I-D.ietf-rtgwg-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 could consider to use [Graceful Shutdown] to facilitate smooth drainage of traffic prior to session tear down, and the Shutdown Communication [[I-D.ietf-idr-shutdown](#)] 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 many network operators are unable to carry out the Voluntary BGP Session Teardown recommendations, because of the operational cost and risk of co-ordinating 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 densely meshed multi-node layer-2 fabric, the caretaker of the fabric could opt to cull BGP sessions on behalf of the stakeholders connected to the fabric.

Such culling of control-plane traffic will pre-empt 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 through the application of a combined layer-3 and layer-4 packet filter deployed in the switched fabric itself.

3.2.1. Packet Filter Considerations

The following considerations apply to the packet filter design:

- o The packet filter MUST only affect BGP traffic specific to the layer-2 fabric, i.e. forming part of the control plane of the system described, rather than multihop BGP traffic which merely transits

- o The packet filter MUST only affect BGP, i.e. TCP/179
- o The packet filter SHOULD make provision for the bidirectional nature of BGP, i.e. that sessions may be established in either direction
- o The packet filter MUST affect all relevant AFIs

[Appendix A](#) contains examples of correct packet filters for various platforms.

3.2.2. Hardware Considerations

Not all hardware is capable of deploying layer 3 / layer 4 filters on layer 2 ports, and even on platforms which support the feature, documented limitations may exist or hardware resource allocation failures may occur during filter deployment which may cause unexpected results. These problems may include:

- o Platform inability to apply layer 3/4 filters on ports which already have layer 2 filters applied
- o Layer 3/4 filters supported for IPv4 but not for IPv6
- o Layer 3/4 filters supported on physical ports, but not on 802.3ad Link Aggregate ports
- o Failure of the operator to apply filters to all 802.3ad Link Aggregate ports
- o Limitations in ACL hardware mechanisms causing filters not to be applied
- o Fragmentation of ACL lookup memory causing transient ACL application problems which are resolved after ACL removal / reapplication
- o Temporary service loss during hardware programming
- o Reduction in hardware ACL capacity if the platform enables lossless ACL application

It is advisable for the operator 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 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 are only deployed for the duration of the maintenance and immediately removed after the maintenance. To prevent unnecessarily 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. 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, and Tore Anderson.

5. Security Considerations

There are no security considerations.

6. IANA Considerations

This document has no actions for IANA.

7. References

7.1. 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, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.

7.2. Informative References

[I-D.ietf-idr-shutdown]
Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", [draft-ietf-idr-shutdown-07](#) (work in progress), March 2017.

[I-D.ietf-rtgwg-bgp-pic]

Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", [draft-ietf-rtgwg-bgp-pic-01](#) (work in progress), June 2016.

[7.3.](#) URIs

[1] <https://github.com/bgp/bgp-session-culling-config-examples>

[Appendix A.](#) Example packet filters

Example packet filters for "Involuntary BGP Session Teardown" at an IXP with LAN prefixes 192.0.2.0/24 and 2001:db8:2::/64.

A repository of configuration examples for a number of assorted platforms can be found at github.com/bgp/bgp-session-culling-config-examples [[1](#)].

[A.1.](#) Cisco IOS, IOS XR & Arista EOS Firewall Example Configuration

```
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.](#) Nokia SR OS Filter Example Configuration


```
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
```

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

