Global Routing Operations
Internet-Draft
Intended status: Best Current Practice
Expires: October 7, 2017

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

April 5, 2017

Mitigating Negative Impact of Maintenance through BGP Session Culling
draft-ietf-grow-bgp-session-culling-00

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 October 7, 2017.

Table of Contents

1. Introduction

In network topologies where BGP speaking routers are directly attached to each other, or use fault detection mechanisms such as BFD [RFC5880], detecting and acting upon a link down event (for example when someone yanks the physical connector) in a timely fashion is straightforward.

However, in topologies where upper layer fast fault detection mechanisms are unavailable and the lower layer topology is hidden from the BGP speakers, operators rely on BGP Hold Timer Expiration (section 6.5 of [RFC4271]) to initiate traffic rerouting. Common BGP Hold Timer values are anywhere between 90 and 180 seconds, which implies a window of 90 to 180 seconds during which traffic blackholing will occur if the lower layer network is not able to

forward traffic.

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.

2. 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.

2.1. 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 would do well to Administratively Shutdown the 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.

2.1.1. Maintenance Communication Considerations

Initiators of the Administrative Shutdown are encouraged to use Shutdown Communication [I-D.ietf-idr-shutdown] to inform the remote side on the nature and duration of the maintenance activities.

2.2. 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.

2.2.1. Packet Filter Considerations

The packet filter should be designed and specified in a way that:

- only affect link-local BGP traffic i.e. forming part of the control plane of the system described, rather than multihop BGP which merely transits

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

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

2.2.2. 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 result. These problems may include:

- Platform inability to apply layer 3/4 filters on ports which 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 802.3ad Link Aggregate ports.
- Failure of the operator to apply filters to all 802.3ad Link Aggregate ports
- Limitations in ACL hardware mechanisms causing filters not to be applied.
- Fragmentation of ACL lookup memory causing transient ACL application problems which 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 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.

2.3. Monitoring 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.

3. Acknowledgments

The authors would like to thank the following people for their contributions to this document: Saku Ytti.

4. Security Considerations

There are no security considerations.

5. IANA Considerations

This document has no actions for IANA.

6. References

6.1. Normative References

[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>.

6.2. 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.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

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.1. Juniper Junos Layer 2 Firewall Example Configuration

```
> show configuration firewall family ethernet-switching filter cull
term towards_peeringlan-v4 {
    from {
        ip-version {
            ipv4 {
                destination-port bgp;
                ip-source-address {
                    192.0.2.0/24;
                }
                ip-destination-address {
                    192.0.2.0/24;
                }
                ip-protocol tcp;
            }
        }
    }
```



```
    then discard;
}
term from_peeringlan-v4 {
    from {
        ip-version {
            ipv4 {
                source-port bgp;
                ip-source-address {
```

```
                    192.0.2.0/24;
                }
                ip-destination-address {
                    192.0.2.0/24;
                }
                ip-protocol tcp;
            }
        }
    }
    then discard;
}
term towards_peeringlan-v6 {
    from {
        ip-version {
            ipv6 {
                next-header tcp;
                destination-port bgp;
                ip6-source-address {
                    2001:db8:2::/64;
                }
                ip6-destination-address {
                    2001:db8:2::/64;
                }
            }
        }
    }
    then discard;
}
term from_peeringlan-v6 {
    from {
        ip-version {
            ipv6 {
                next-header tcp;
                source-port bgp;
                ip6-source-address {
                    2001:db8:2::/64;
                }
                ip6-destination-address {
                    2001:db8:2::/64;
                }
```



```
            }
```

```
            }
        }
        then discard;
    }
    term rest {
        then accept;
    }

    > show configuration interfaces xe-0/0/46
    description "IXP participant affected by maintenance"
    unit 0 {
        family ethernet-switching {
            filter {
                input cull;
            }
        }
    }
```

A.2. 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
    !
```

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