Inter-Domain Routing
Internet-Draft
Updates: 4271 (if approved)
Intended status: Standards Track
Expires: August 22, 2021

M. Aelmans
Juniper Networks
M. Stucchi
Independent
J. Snijders
Fastly

February 18, 2021

Revised BGP Maximum Prefix Limits Outbound
draft-sas-idr-maxprefix-outbound-02

Abstract

This document updates RFC4271 by adding a control mechanism which limits the negative impact of outbound route leaks (RFC7908) in order to prevent resource exhaustion in Border Gateway Protocol (BGP) implementations.

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.

Status of This Memo

This Internet-Draft will expire on August 22, 2021.

Copyright Notice

Table of Contents

1. Introduction

This document updates [RFC4271] by adding a control mechanism which limits the negative impact of outbound route leaks [RFC7908] in order to prevent resource exhaustion in Border Gateway Protocol (BGP) implementations. [RFC4271] describes methods to tear down BGP sessions or discard UPDATES after certain inbound thresholds are exceeded. In addition to "inbound maximum prefix limits", this document introduces a specification for "outbound maximum prefix limits". [I-D.sas-idr-maxprefix-inbound] updates sections in [RFC4271] to clarify "inbound maximum prefix limits". This documents updates those sections again to add "outbound maximum prefix limits".

2. Changes to RFC4271 Section 6

This section updates [RFC4271] to specify what events can result in AutomaticStop (Event 8) in the BGP FSM.

The following paragraph replaces the second paragraph of Section 6.7 (Cease), which starts with "A BGP speaker MAY support" and ends with "The speaker MAY also log this locally.":

> A BGP speaker MAY support the ability to impose a locally-configured, upper bound on the number of address prefixes the speaker is willing to accept from a neighbor (inbound maximum prefix limit) or send to a neighbor (outbound prefix limit). The limit on the prefixes accepted from a neighbor can be applied before policy processing (Pre-Policy) or after policy processing (Post-Policy). Outbound prefix limits MUST be measured after policy, since the Policy (even a policy of "send all") is run before determining what can be sent. When the upper bound is reached, the speaker, under control of local configuration, either:
>
> A. Discards new address prefixes being sent to the neighbor while maintaining the BGP connection with the neighbor. As these prefixes are discared and their reachability information is not sent to the neighbor it might lead to inconsistent routing behaviour;
>
> B. Sent all prefixes exceeding the threshold and generates a log;
>
> C. Terminates the BGP session with the neighbor. This should be done using a Hard Reset according to [RFC8538].
>
> If the BGP speaker uses option (b), where the limit causes a CEASE Notification, then the CEASE error codes should use:

Subcode	Symbolic Name
1	Threshold exceeded: Maximum Number of Prefixes Received
TBD	Threshold exceeded: Maximum Number of Prefixes Sent

> The speaker MAY also log the event locally.

3. Changes to RFC4271 Section 8

This section updates Section 8 [RFC4271], the paragraph that starts with "One reason for an AutomaticStop event is" and ends with "The local system automatically disconnects the peer." is replaced with:

> Possible reasons for an AutomaticStop event are: A BGP speaker receives an UPDATE messages with a number of prefixes for a given peer such that the total prefixes received exceeds the maximum number of prefixes configured (either "Pre-Policy" or "Post-Policy"), or announces more prefixes than through local configuration allowed to. The local system automatically disconnects the peer.

4. Changes to RFC4271 Section 9

This section updates [RFC4271] by adding a subsection after Section 9.4 (Originating BGP routes) to specify various events that can lead up to an AutomaticStop (Event 8) in the BGP FSM.

9.5 Maximum Prefix Limits

9.5.1 Pre-Policy Inbound Maximum Prefix Limits

The Adj-RIBs-In stores routing information learned from inbound UPDATE messages that were received from another BGP speaker Section 3.2 [RFC4271]. The pre-policy limit uses the number of NLRIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI) as input into its threshold comparisons. For example, when an operator configures the pre-policy limit for IPv4 Unicast to be 50 on a given EBGP session, and the other BGP speaker announces its 51st IPv4 Unicast NLRI, the session MUST be terminated.

Pre-policy limits are particularly useful to help dampen the effects of full table route leaks and memory exhaustion when the implementation stores rejected routes.

9.5.2 Post-Policy Inbound Maximum Prefix Limits

[RFC4271] describes a Policy Information Base (PIB) that contains local policies that can be applied to the information in the Routing Information Base (RIB). The post-policy limit uses the number of NLRIs per Address Family Identifier (AFI)

per Subsequent Address Family Identifier (SAFI), after application of the Import Policy as input into its threshold comparisons. For example, when an operator configures the post-policy limit for IPv4 Unicast to be 50 on a given EBGP session, and the other BGP speaker announces a hundred IPv4 Unicast routes of which none are accepted as a result of the local import policy (and thus not considered for the Loc-RIB by the local BGP speaker), the session is not terminated.

Post-policy limits are useful to help prevent FIB exhaustion and prevent accidental BGP session teardown due to prefixes not accepted by policy anyway.

9.5.3 Outbound Maximum Prefix Limits

An operator MAY configure a BGP speaker to terminate its BGP session with a neighbor when the number of address prefixes to be advertised to that neighbor exceeds a locally configured post-policy upper limit. The BGP speaker then MUST send the neighbor a NOTIFICATION message with the Error Code "Cease" and the Error Subcode "Threshold reached: Maximum Number of Prefixes Sent". Implementations MAY support additional actions. The Hard Cease action is defined in [RFC8538].

Reporting when thresholds have been exceeded is an implementation specific consideration, but SHOULD include methods such as Syslog [RFC5424]. By definition, Outbound Maximum Prefix Limits are Post-Policy.

The Adj-RIBs-Out stores information selected by the local BGP speaker for advertisement to its neighbors. The routing information stored in the Adj-RIBs-Out will be carried in the local BGP speaker's UPDATE messages and advertised to its neighbors Section 3.2 [RFC4271]. The Outbound Maximum Prefix Limit uses the number of NLRIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI), after application of the Export Policy, as input into its threshold comparisons. For example, when an operator configures the Outbound Maximum Prefix Limit for IPv4 Unicast to be 50 on a given EBGP session, and were about to announce its 51st IPv4 Unicast NLRI to the other BGP speaker as a result of the local export policy, the session MUST be terminated.

Outbound Maximum Prefix Limits are useful to help dampen the negative effects of a misconfiguration in local policy. In

many cases, it would be more desirable to tear down a BGP session rather than causing or propagating a route leak.

5. Use cases

Egress maximum prefix limits are usefull in a variety of cases. Some of those are outlined in this section.

5.1. Internet use case

In order to prevent the BGP speaker from leaking a full routing table to its neighbor operators should implement proper routing policy and preferably RFC8212. However, even when implementing both measurements an operator could still (accidentaly) announce more routes than intended. Setting a maximum prefix outbound value prevents this.

5.2. CE protection

Residential and many business customers connected to the internet using a 'simple' CPE and connected to a single Service Provider only needs to accept a single default route and not the full internet table. In order to prevent overloading the CPE Control Plane, maximum outbound limits should be applied on the session on the PE router.

5.3. PE-CE BGP session from operator side

-- Change this so it explains that it's extra protection towards the PE so it won't kill the BGP session due to max prefix inbound -- Internet providers PE side gateway PE-CE connections would would generally set maximum prefix to disconnect if maximum prefix is reached. This is a secondary protection mechanism as the primary is prefix length and AS path checks.

6. Security Considerations

Maximum Prefix Limits are an essential tool for routing operations and SHOULD be used to increase stability. They provide a first-line mechanism to avoid route leaks and to avoid unintended routing suggestions to happen between neighbors. Implementing this measures is only one of the building blocks you need to provide full security, but it is important to build a modular defense system.

Stability for the routing table is also an important aspect for implementing the measures included in this draft. Ensuring that neighbors will not receive an amount of routes that would overload

their routing platform contributes to the stability of interconnections and of the Internet as a whole.

7. IANA Considerations

This memo requests that IANA assigns a new subcode named "Threshold exceeded: Maximum Number of Prefixes Sent" in the "Cease NOTIFICATION message subcodes" registry under the "Border Gateway Protocol (BGP) Parameters" group.

8. Acknowledgments

The authors would like to thank Saku Ytti and John Heasley (NTT Ltd.), Jeff Haas, Colby Barth and John Scudder (Juniper Networks), Martijn Schmidt (i3D.net), Teun Vink (BIT), Sabri Berisha (eBay), Martin Pels (Quanza), Steven Bakker (AMS-IX), Aftab Siddiqui (ISOC), Yu Tianpeng, Ruediger Volk (Deutsche Telekom), Robert Raszuk (Bloomberg), Jakob Heitz (Cisco), Warren Kumari (Google), Ben Maddison (Workonline), Randy Bush, Brian Dickson and Gyan Mishra (Verizon) for their support, insightful reviews, and comments.

9. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in RFC7942. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

The table below provides an overview (as of the moment of writing) of which vendors have produced implementations of inbound or outbound maximum prefix limits. Each table cell shows the applicable configuration keywords if the vendor implemented the feature.

Vendor	Inbound Pre-Policy	Inbound Post-Policy	Outbound
Cisco IOS XR		maximum-prefix	
Cisco IOS XE		maximum-prefix	
Juniper Junos OS	prefix-limit	accepted-prefix-limit, or prefix-limit combined with 'keep none'	advertise-prefix-limit *
Nokia SR OS	prefix-limit		
NIC.CZ BIRD	'import keep filtered' combined with 'receive limit'	'import limit' or 'receive limit'	export limit
OpenBSD OpenBGPD	max-prefix		
Arista EOS	maximum-routes	maximum-accepted-routes	
Huawei VRPv5	peer route-limit		
Huawei VRPv8	peer route-limit	peer route-limit accept-prefix	

First presented by Job Snijders at [RIPE77]

Table 1: Maximum prefix limits capabilities per implementation

*In testing stage

10. References

10.1. Normative References

[I-D.sas-idr-maxprefix-inbound]
Aelmans, M., stucchi-lists@glevia.com, s., and J. Snijders, "BGP Maximum Prefix Limits Inbound", draft-sas-idr-maxprefix-inbound-01 (work in progress), October 2020.

[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>.

[RFC8538] Patel, K., Fernando, R., Scudder, J., and J. Haas, "Notification Message Support for BGP Graceful Restart", RFC 8538, DOI 10.17487/RFC8538, March 2019, <https://www.rfc-editor.org/info/rfc8538>.

10.2. Informative References

[I-D.ietf-idr-bgp-model]
Jethanandani, M., Patel, K., Hares, S., and J. Haas, "BGP YANG Model for Service Provider Networks", draft-ietf-idr-bgp-model-10 (work in progress), November 2020.

[RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, DOI 10.17487/RFC5424, March 2009, <https://www.rfc-editor.org/info/rfc5424>.

[RFC7908] Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", RFC 7908, DOI 10.17487/RFC7908, June 2016, <https://www.rfc-editor.org/info/rfc7908>.

[RIPE77] Snijders, J., "Robust Routing Policy Architecture", May 2018, <https://ripe77.ripe.net/wp-content/uploads/presentations/59-RIPE77_Snijders_Routing_Policy_Architecture.pdf>.

Authors' Addresses

Melchior Aelmans
Juniper Networks
Boeing Avenue 240
Schiphol-Rijk 1119 PZ
Netherlands

Email: maelmans@juniper.net

Massimiliano Stucchi
Independent

Email: max@stucchi.ch

Job Snijders
Fastly
Amsterdam
Netherlands

Email: job@fastly.com