Internet Engineering Task Force
Internet-Draft
Intended status: BCP
Expires: September 3, 2012

J. Durand
CISCO Systems, Inc.
I. Pepelnjak
NIL
G. Doering
SpaceNet

March 2, 2012

BGP operations and security
draft-jdurand-bgp-security-00.txt

Abstract

This documents describes best current practices to manage securely BGP in a network. It will explain the basic policies ones should configure on BGP peerings to keep an healthy BGP table. This document will only focus on unicast and multicast tables (SAFI 1 and 2) for IPv4 and IPv6.

Foreword

A placeholder to list general observations about this document.

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 [1].

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

BGP [6] is the protocol used in the internet to exchange routing information between network domains. This protocol does not directly include mechanisms that control that routes exchanged conform to the various rules defined by the Internet community. This document intends to summarize most common existing rules and help network administrators applying simply coherent BGP policies.

2. Definitions

- BGP peering: any TCP BGP connection on the Internet.

3. Protection of BGP sessions

3.1. MD5 passwords on BGP peerings

BGP sessions can be secured with MD5 passwords [8], to protect against attacks that could bring down the session (by sending spoofed TCP RST packets) or possibly insert packets into the TCP stream (routing attacks).

The drawback of TCP/MD5 is additional management overhead for password maintenance. MD5 protection is recommended when peerings are established over shared networks where spoofing can be done (like internet exchanges, IXPs).

You should block spoofed packets (packets with source IP address belonging to your IP address space) at all edges of your network, making TCP/MD5 protection of BGP sessions unnecessary on iBGP session or EBGP sessions run over point-to-point links.

3.2. BGP TTL security

BGP sessions can be made harder to spoof with the TTL security - instead of sending TCP packets with TTL value = 1, the routers send the TCP packets with TTL value = 255 and the receiver checks that the TTL value equals 255. Since it's impossible to send an IP packet with TTL = 255 to a non-directly-connected IP host, BGP TTL security effectively prevents all spoofing attacks coming from third parties not directly connected to the same subnet as the BGP-speaking routers.

Note: Like MD5 protection, TTL security has to be configured on both ends of a BGP session.

4. Prefix filtering

The main aspect of securing BGP resides in controlling the prefixes that are received/advertised on the BGP peerings. Prefixes exchanged between BGP peers are controlled with inbound and outbound filters that can match on IP prefixes (prefix filters, Section 4), AS paths (as-path filters, Section 7) or any other attributes of a BGP prefix (for example, BGP communities, Section 8).

4.1. Definition of prefix filters

This section list the most commonly used prefix filters. Following sections will clarify where these filters should be applied.

4.1.1. Prefixes that are not routable by definition

4.1.1.1. IPv4

RFC3330 [12] clarifies "special" IPv4 prefixes and their status in the Internet. Following prefixes MUST NOT cross network boundaries (ie. ASN) and therefore MUST be filtered:

- 10.0.0.0/8 and more specific - private use

- 169.254.0.0/16 and more specific - link-local

- 172.0.0.0/8 and more specific - loopbacks

- 172.16.0.0/12 and more specific - private use

- 192.0.2.0/24 and more specific- documentation

- 192.168.0.0/16 and more specific - private use

- 224.0.0.0/4 and more specific - multicast

- 240.0.0.0/4 and more specific - reserved

4.1.1.2. IPv6

There is no equivalent of RFC3300 for IPv6. This document recalls the prefixes that MUST not cross network boundaries and therefore MUST be filtered:

- 2001:DB8::/32 and more specific - documentation [13]

- Prefixes more specific than 2002::/16 - 6to4 [3]

- 3FFE::/16 and more specific - was initially used for the 6Bone (worldwide IPv6 test network) and returned to IANA.

- FC00::/7 and more specific - ULA (Unique Local Addresses) [5]

- FE80::/10 and more specific - link-local addresses [7]

- FEC0::/10 and more specific - initially reserved for unicast site-local addresses [4]. As some networks may still use it for private addressing it is worth considering it when filtering private prefixes.

- FF00::/8 and more specific - multicast

The list of IPv6 prefixes that MUST not cross network boundaries can be simplified as follows:

- 2001:DB8::/32 and more specific - documentation [13]

- Prefixes more specific than 2002::/16 - 6to4 [3]

- All prefixes that are outside 2000::/3 prefix

4.1.2. Prefixes not allocated

IANA allocates prefixes to RIRs which in turn allocate prefixes to LIRs. It is wise not to accept in the routing table prefixes that are not allocated. This could mean allocation made by IANA and/or allocations done by RIRs. This section details the options for building list of allocated prefixes at every level.

4.1.2.1. IANA allocated prefixes filters

IANA has allocated all the IPv4 available space. Therefore there is no reason why one would keep checking prefixes are in the IANA allocated address space [19]. No specific filter need to be put in place by administrators who want to make sure that IPv4 prefixes they receive have been allocated by IANA.

For IPv6, given the size of the address space, it can be seen as wise accepting only prefixes derived from those allocated by IANA. Administrators can dynamically build this list from the IANA allocated IPv6 space [20]. As IANA keeps allocating prefixes to RIRs, the aforementioned list should be checked regularly against changes and if they occur, prefix filter should be computed and pushed on network devices. As there is delay between the time a RIR receives a new prefix and the moment it starts allocating portions of it to its LIRs, there is no need doing this step quickly and

frequently. At least process in place should make sure there is no more than one month between the time the IANA IPv6 allocated prefix list changes and the moment all IPv6 prefix filters have been updated.

4.1.2.2. RIR allocated prefixes filters

A more precise check can be performed as one would like to make sure that prefixes they receive are being originated by the autonomous system which actually own the prefix. It has been observed in the past that one could easily advertise someone else's prefix (or more specific prefixes) and create black holes or security threats. To overcome that risk, administrators would need to make sure BGP advertisements correspond to information located in the existing registries. At this stage 2 options can be considered (short and long term options). They are described in the following subsections.

4.1.2.3. Prefix filters creation from RIR database

This option consists in using RIR database information for building for a given BGP neighbor a list of prefixes and the list of prefix with corresponding originating autonomous system. This can be done relatively easily using scripts and existing tools capable of retrieving this information in the registries. This approach is exactly the same for both IPv4 and IPv6.

The macro-algorithm for the script is described as follows. For the peer that is considered, the distant network administrator has provided the autonomous system and may be able to provide an AS-SET object (aka AS-MACRO). An AS-SET is an object which contains AS numbers or other AS-SET's. An operator may create an AS-SET defining all the AS numbers of its customers. A tier 1 transit provider might create an AS-SET describing the AS-SET of connected operators, which in turn describe the AS numbers of their customers. Using recursion, it is possible to retrieve from an AS-SET the complete list of AS numbers that the peer is susceptible to announce. For each of these AS numbers, it is also easy to check in the corresponding RIR database all associated prefixes. With these 2 mechanisms a script can build for a given peer the list of allowed prefixes and the AS number from which they should be originated.

As prefixes, AS numbers and AS-SET's may not all be under the same RIR authority, a difficulty resides choosing for each object the appropriate database to poll. Some registries have been created and are not restricted to a given region or authoritative RIR. They allow RIRs to publish their information in a common place. They also make it possible for any subscriber (probably under contract) to publish information too. When doing requests inside such a database,

it is possible to specify the source of information in order to have the most reliable data. One could check the central registry and only check that the source is one of the 5 RIRs. The probably most famous registry of that kind is the RADB [21] (Routing Assets Database).

As objects in RIRs DB may quickly vary over time, it is important that prefix filters computed using this mechanism are refreshed regularly. A daily basis could even been considered as some routing changes must be done sometimes in a certain emergency and registries may be updated at the very last moment. It has to be noted that this approach significantly increases the complexity of the router configurations as it can quickly add more than ten thousands configuration lines for some important peers.

4.1.2.4. SIDR - Secure Inter Domain Routing

IETF has created a working group called SIDR (Secure Inter-Domain Routing) in order to create an architecture to secure internet advertisements. At the time this document is written, many document has been published and a framework is proposed so that advertisements can be checked against signed routing objects in RIR routing registries. Implementing mechanisms proposed by this working group is the solution that will solve at a longer term the BGP routing security. But as it may take time objects are signed and deployments are done such a solution will need to be combined at the time being with other mechanisms proposed in this document. The rest of this section assumes the reader understands all technologies associated with SIDR.

Each received route on a router should be checked against the RPKI data set: if a corresponding ROA is found and is valid then the prefix should be accepted. It the ROA is found and is INVALID then the prefix should be discarded. If an ROA is not found then the prefix should be accepted but corresponding route should be given a low preference.

4.1.3. Prefixes too specific

4.1.3.1. IPv4

Prefixes longer than /24 are usually not announced in the IPv4 internet [16]

4.1.3.2. IPv6

Prefixes longer than /48 are usually not announced in the IPv6 internet [17]

4.1.4. Anti-spoofing filters

Filtering its own prefixes on peerings with all peers (ingress direction) is a protection against spoofing attacks. Such filters must be defined with caution as they can break existing redundancy mechanisms. For example in case an operator has a multihomed customer, it should keep accepting the customer prefix from its peers and upstreams. This will make it possible for the customer to keep accessing its operator network (and other customers) via the internet in case the BGP peering between the customer and the operator is down.

4.1.5. Exchange point LAN prefixes

When a network is present on an exchange point, it must make sure it doesn't receive exchange point LAN prefix and more specifics from any of its BGP peers.

4.1.6. Default route

4.1.6.1. IPv4

0.0.0.0/0 prefix MUST NOT be announced on the Internet but it is usually exchanged on upstream/customer peerings.

4.1.6.2. IPv6

::/0 prefix MUST NOT be announced on the Internet but it is usually exchanged on upstream/customer peerings.

4.2. Prefix filtering recommendations in full routing networks

For networks that have the full internet BGP table, some policies should be applied on each BGP peer for received and advertised routes. It is recommended that each autonomous filter configures rules for advertised and received routes at all its borders as this will protect the network and its peer even in case of misconfiguration. The most commonly used filtering policy is proposed in this section.

4.2.1. Filters with internet peers

4.2.1.1. Ingress filtering

There are basically 2 options, the loose one where no check will be done against RIR allocations and the strict one where it will be verified that announcements strictly conform to what is declared in routing registries.

4.2.1.1.1. Ingress filtering loose option

In that case, the following prefixes received from a BGP peer will be filtered:

- Prefixes not routable (Section 4.1.1)

- Prefixes not allocated by IANA (IPv6 only) (Section 4.1.2.1)

- Routes too specific (Section 4.1.3)

- Self prefixes (Section 4.1.4)

- Exchange points LAN prefixes (Section 4.1.5)

- Default route (Section 4.1.6)

4.2.1.1.2. Ingress filtering strict option

In that case, filters are applied to make sure advertisements strictly conform to what is declared in routing registries Section 4.1.2.2. It must be checked that in case of script failure all routes are rejected.

In addition to this, one could apply following filters beforehand in case routing registry used as source of information by the script is not fully trusted:

- Prefixes not routable (Section 4.1.1)

- Routes too specific (Section 4.1.3)

- Self prefixes (Section 4.1.4)

- Exchange points LAN prefixes (Section 4.1.5)

- Default route (Section 4.1.6)

4.2.1.2. Egress filtering

Configuration in place will make sure that only appropriate prefixes are sent. These can be for example prefixes belonging to the considered networks and those of its customers. This can be done using BGP communities or many other solution. Whatever scenario considered, it can be desirable that following filters are positioned before to avoid unwanted route announcement due to bad configuration:

- Prefixes not routable (Section 4.1.1)

- Routes too specific (Section 4.1.3)

- Exchange points LAN prefixes (Section 4.1.5)

- Default route (Section 4.1.6)

In case it is possible to list the prefixes to be advertised, then just configuring the list of allowed prefixes and denying the rest is sufficient.

4.2.2. Filters with customers

4.2.2.1. Ingress filtering

Ingress policy with end customers is pretty straightforward: only customers prefixes must be accepted, all others should be discarded. The list of accepted prefixes can be manually specified, after having verified that they are valid. This validation can be done with the appropriate IP address management authorities. For example one will not accept a prefix if it is in a PA (Provider Aggregateable) block.

Same rules apply in case the customer is also a network connecting other customers (for example a tier 1 transit provider connecting service providers). An exception can be envisaged in case it is known that the customer network applies strict ingress/egress filtering, and the number of prefixes announced by that network is too large to list them in the router configuration. In that case filters as in Section 4.2.1.1 can be applied.

4.2.2.2. Egress filtering

Egress policy with customers may vary according to the routes customer wants to receive. In the simplest possible scenario, customer wants to receive only the default route, which can be done easily by applying a filter with the default route only.

In case the customer wants to receive the full routing (in case it is multihomed or if wants to have a view on the internet table), the following filters can be simply applied on the BGP peering:

- Prefixes not routable (Section 4.1.1)

- Routes too specific (Section 4.1.3)

- Default route (Section 4.1.6)

There can be a difference for the default route that can be announced to the customer in addition to the full BGP table. This can be done simply by removing the filter for the default route. As the default route may not be present in the routing table, one may decide to originate it only for peerings where it has to be advertised.

4.2.3. Filters with upstream providers

4.2.3.1. Ingress filtering

In case the full routing table is desired from the upstream, the prefix filtering to apply is more or less the same than the one for peers Section 4.2.1.1. There can be a difference for the default route that can be desired from an upstream provider even if it advertises the full BGP table. In case the upstream provider is supposed to announce only the default route, a simple filter will be applied to accept only the default prefix and nothing else.

4.2.3.2. Egress filtering

The filters to be applied should not differ from the ones applied for internet peers (Section 4.2.1.2).

4.3. Prefix filtering recommendations for leaf networks

4.3.1. Ingress filtering

The leaf network will position the filters corresponding to the routes it is requesting from its upstream. In case a default route is requested, simple inbound filter will be applied to accept only that default route (Section 4.1.6). In case the leaf network is not capable of listing the prefix because the amount is too large (for example if it requires the full internet routing table) then it should configure filters to avoid receiving bad announcements from its upstream:

- Prefixes not routable (Section 4.1.1)

- Routes too specific (Section 4.1.3)

- Self prefixes (Section 4.1.4)

- Default route (Section 4.1.6) depending if the route is requested or not

4.3.2. Egress filtering

A leaf network will most likely have a very straightforward policy: it will only announce its local routes. It can also configure the following prefixes filters described in Section 4.2.1.2 to avoid announcing invalid routes to its upstream provider.

5. BGP route flap dampening

BGP route flap dampening mechanism makes it possible to give penalties to routes each time they change in the BGP routing table. Initially this mechanism was created to protect the entire internet from multiple events impacting a single network. RIPE community now recommends not using BGP route flap dampening [15]. Author of this document proposes to follow the proposal of the RIPE community.

6. Maximum prefixes on a peering

It is recommended to configure a limit on the number of routes to be accepted from a peer. Following rules are generally recommended:

- From peers, it is recommended to have a limit lower than the number of routes in the internet. This will shut down the BGP peering if the peer suddenly advertises the full table. One can also configure different limits for each peer, according to the number of routes they are supposed to advertise.

- From upstreams which provide full routing, it is recommended to have a limit much higher than the number of routes in the internet. A limit is still useful in order to protect the network (and in particular the routers' memory) if too many routes are sent by the upstream. The limit should be chosen according to the number of routes that can actually be handled by routers.

It is important to review regularly the limits that are configured as the internet can quickly change over time. Some vendors propose mechanisms to have 2 thresholds: while the higher number specified will shutdown the peering, the first threshold will only trigger a log and can be used to passively adjust limits based on observations made on the network.

7. AS-path filtering

The following rules should be applied on BGP AS-paths:

- Do not accept anything other than customer's AS number from the customer. Alternatively, only accept AS-paths with a single AS number (potentially repeated several times) from your customers. The latter option is easier to configure than per-customer AS-path filters: the default BGP logic will make sure in that case that the first AS number in the AS-path is the one of the peer.

- Do not accept overly long AS path prepending from the customer.

- Do not accept more than two distinct AS path numbers in the AS path if your customer is an ISP with customers. This rule becomes useless in case prefix filters are built from registries as described in Section 4.1.2.3.

- Do not advertise prefixes with non-empty AS-path if you're not transit.

- Do not advertise prefixes with upstream AS numbers in the AS path to your peering AS.

- Do not accept private AS numbers except from customers

- Do not advertise private AS numbers. Exception: Customers using BGP without having their own AS number must use private AS numbers to advertise their prefixes to their upstream. The private AS number is usually provided by the upstream.

- Do not accept prefixes when the first AS number in the AS-path is not the one of the peer. In case the peering is done toward a BGP route-server [23] (connection on an Internet eXchange Point - IXP) with transparent AS path handling, this verification needs to be de-activated as the first AS number will be the one of an IXP member whereas the peer AS number will be the one of the BGP route-server.

8. BGP community scrubbing

Optionally we can consider the following rules on BGP AS-paths:

- Scrub inbound communities with your AS number in the high-order bits - allow only those communities that customers/peers can use as a signaling mechanism

- Do not remove other communities: your customers might need them to communicate with upstream providers. In particular do not (generally) remove the no-export community as it is usually announced by your peer for a certain purpose.

9. Acknowledgements

A placeholder to acknowledge contributors.

10. IANA Considerations

This memo includes no request to IANA.

11. Security Considerations

This document is entirely about BGP operational security.

12. References

12.1. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://xml.resource.org/public/rfc/html/rfc2119.html>.

[2] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

[3] Carpenter, B. and K. Moore, "Connection of IPv6 Domains via IPv4 Clouds", RFC 3056, February 2001.

[4] Huitema, C. and B. Carpenter, "Deprecating Site Local Addresses", RFC 3879, September 2004.

[5] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

[6] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[7] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[8] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

12.2. Informative References

[9] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5,

RFC 1918, February 1996.

[10] Crocker, D., Ed. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, November 1997.

[11] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[12] IANA, "Special-Use IPv4 Addresses", RFC 3330, September 2002.

[13] Huston, G., Lord, A., and P. Smith, "IPv6 Address Prefix Reserved for Documentation", RFC 3849, July 2004.

[14] Crocker, D., Ed. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", RFC 4234, October 2005.

[15] Smith, P. and C. Panigl, "RIPE-378 - RIPE Routing Working Group Recommendations On Route-flap Damping", May 2006.

[16] Smith, P., Evans, R., and M. Hughes, "RIPE-399 - RIPE Routing Working Group Recommendations on Route Aggregation", December 2006.

[17] Smith, P. and R. Evans, "RIPE-532 - RIPE Routing Working Group Recommendations on IPv6 Route Aggregation", November 2011.

[18] Doering, G., "IPv6 BGP Filter Recommendations", November 2009, <http://www.space.net/~gert/RIPE/ipv6-filters.html>.

[19] "IANA IPv4 Address Space Registry", <http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml>.

[20] "IANA IPv6 Address Space Registry", <http://www.iana.org/assignments/ipv6-unicast-address-assignments/ipv6-unicast-address-assignments.xml>.

[21] "Routing Assets Database", <http://www.radb.net>.

[22] "Secure Inter-Domain Routing IETF working group", <http://datatracker.ietf.org/wg/sidr/>.

[23] "Internet Exchange Route Server", <http://tools.ietf.org/id/draft-jasinska-ix-bgp-route-server-03.txt>.

Authors' Addresses

Jerome Durand
CISCO Systems, Inc.
11 rue Camille Desmoulins
Issy-les-Moulineaux 92782 CEDEX
FR

Email: jerduran@cisco.com

Ivan Pepelnjak
NIL Data Communications
Tivolska 48
Ljubljana 1000
Slovenia

Email: ip@nil.com

Gert Doering
SpaceNet AG
Joseph-Dollinger-Bogen 14
Muenchen D-80807
Germany

Email: gert@space.net