Internet Engineering Task Force
Internet-Draft
Intended status: BCP
Expires: March 25, 2013

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

September 21, 2012

BGP operations and security
draft-jdurand-bgp-security-02.txt

Abstract

BGP (Border Gateway Protocol) is the protocol almost exclusively used in the Internet to exchange routing information between network domains. Due to this central nature, it's important to understand the security measures that can and should be deployed to prevent accidental or intentional routing disturbances.

This document describes measures to protect the BGP sessions itself (like TTL, MD5, control plane filtering) and to better control the flow of routing information, using prefix filtering and automatization of prefix filters, max-prefix filtering, AS path filtering, route flap dampening and BGP community scrubbing.

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 . . . . . . . . . . . . . . . . . . . . . . . . .  4
   2.  Definitions  . . . . . . . . . . . . . . . . . . . . . . . . .  4
   3.  Protection of BGP router . . . . . . . . . . . . . . . . . . .  4
   4.  Protection of BGP sessions . . . . . . . . . . . . . . . . . .  4
     4.1.  Protection of TCP sessions used by BGP . . . . . . . . . .  4
     4.2.  BGP TTL security . . . . . . . . . . . . . . . . . . . . .  5
   5.  Prefix filtering . . . . . . . . . . . . . . . . . . . . . . .  5
     5.1.  Definition of prefix filters . . . . . . . . . . . . . . .  5
       5.1.1.  Prefixes that MUST not be routed by definition . . . .  5
       5.1.2.  Prefixes not allocated . . . . . . . . . . . . . . . .  6
       5.1.3.  Prefixes too specific  . . . . . . . . . . . . . . . .  9
       5.1.4.  Filtering prefixes belonging to local AS . . . . . . .  9
       5.1.5.  Internet exchange point (IXP) LAN prefixes . . . . . . 10
       5.1.6.  Default route  . . . . . . . . . . . . . . . . . . . . 11
     5.2.  Prefix filtering recommendations in full routing
           networks . . . . . . . . . . . . . . . . . . . . . . . . . 11
       5.2.1.  Filters with internet peers  . . . . . . . . . . . . . 12
       5.2.2.  Filters with customers . . . . . . . . . . . . . . . . 13
       5.2.3.  Filters with upstream providers  . . . . . . . . . . . 14
     5.3.  Prefix filtering recommendations for leaf networks . . . . 14
       5.3.1.  Inbound filtering  . . . . . . . . . . . . . . . . . . 14
       5.3.2.  Outbound filtering . . . . . . . . . . . . . . . . . . 15
   6.  BGP route flap dampening . . . . . . . . . . . . . . . . . . . 15
   7.  Maximum prefixes on a peering  . . . . . . . . . . . . . . . . 15
   8.  AS-path filtering  . . . . . . . . . . . . . . . . . . . . . . 16
   9.  Next-Hop Filtering . . . . . . . . . . . . . . . . . . . . . . 17
   10. BGP community scrubbing  . . . . . . . . . . . . . . . . . . . 17
   11. Change logs  . . . . . . . . . . . . . . . . . . . . . . . . . 18
     11.1. Diffs between draft-jdurand-bgp-security-01 and
           draft-jdurand-bgp-security-00  . . . . . . . . . . . . . . 18
     11.2. Diffs between draft-jdurand-bgp-security-02 and
           draft-jdurand-bgp-security-01  . . . . . . . . . . . . . . 19
   12. Acknowledgements . . . . . . . . . . . . . . . . . . . . . . . 19
   13. IANA Considerations  . . . . . . . . . . . . . . . . . . . . . 20
   14. Security Considerations  . . . . . . . . . . . . . . . . . . . 20
   15. References . . . . . . . . . . . . . . . . . . . . . . . . . . 20
     15.1. Normative References . . . . . . . . . . . . . . . . . . . 20
     15.2. Informative References . . . . . . . . . . . . . . . . . . 21
   Authors' Addresses . . . . . . . . . . . . . . . . . . . . . . . . 22
```

1. Introduction

BGP [7] 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 router

The BGP router needs to be protected from stray packets. This protection should be achieved by an access-list (ACL) which would discard all packets directed to TCP port 179 on the local device and sourced from an address not known to be a BGP neighbor. If supported, an ACL specific to the control-plane of the router should be used (receive-ACL, control-plane policing, etc.), to avoid filtering transit traffic if not needed. If the hardware can not do that, interface ACLs can be used to block packets to the local router.

Some routers automatically program such an ACL upon BGP configuration. On other devices this ACL should be configured and maintained manually or using scripts.

The filtering of packets destined to the local router is a wider topic than "just for BGP" (if you bring down a router by overloading one of the other protocols from remote, BGP is harmed as well). For a more detailed recommendation, see RFC6192 [19].

4. Protection of BGP sessions

4.1. Protection of TCP sessions used by BGP

Attacks on TCP sessions used by BGP (ex: sending spoofed TCP RST packets) could bring down the TCP session. Following a successful ARP spoofing attack (or other similar Man-in-the-Middle attack), the attacker might even be able to inject packets into the TCP stream (routing attacks).

TCP sessions used by BGP can be secured with a variety of mechanisms.

MD5 protection of TCP session header [2] is the most common one, but one could also use IPsec or TCP Authentication Option (TCP-AO, [10]).

The drawback of TCP session protection is additional configuration and management overhead for authentication information (ex: MD5 password) maintenance. Protection of TCP sessions used by BGP is thus 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 the protection of TCP sessions used by BGP unnecessary on iBGP session or EBGP sessions run over point-to-point links.

4.2. BGP TTL security

BGP sessions can be made harder to spoof with the TTL security [9]. 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.

5. 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 5), AS paths (as-path filters, Section 8) or any other attributes of a BGP prefix (for example, BGP communities, Section 10).

5.1. Definition of prefix filters

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

5.1.1. Prefixes that MUST not be routed by definition

5.1.1.1. IPv4

At the time of the writing of this document, there is no dynamic IPv4 registry listing special prefixes and their status on the internet. On the other hand static document RFC5735 [17] clarifies "special" IPv4 prefixes and their status in the Internet. Since publication of that RFC another prefix has been added on the list of the special use prefixes. Following prefixes MUST NOT cross network boundaries (ie. ASN) and therefore MUST be filtered:

- Prefixes defined in RFC5735 [17] and more specifics

- Shared address space [31] - 100.64.0.0/10 and more specifics

5.1.1.2. IPv6

IPv6 registry [26] maintains the list of IPv6 special purpose prefixes. With the exception of the 6to4 2002::/16 prefix in that registry, all other prefixes that are mentioned and more specifics MUST not cross network boundaries and therefore MUST be filtered. The 6to4 prefix 2002::/16 is an exception because the prefix itself can be advertised, but more specifics MUST be filtered according to [4], section 5.2.3.

At the time of the writing of this document, the list of IPv6 prefixes that MUST not cross network boundaries can be simplified as IANA allocates at the time being prefixes to RIR's only in 2000::/3 prefix [25]. All other prefixes (ULA's, link-local, multicast... are outside of that prefix) and therefore the simplified list becomes:

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

- Prefixes more specifics than 2002::/16 - 6to4 [4]

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

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

5.1.2. 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. It is important to understand that filtering prefixes not allocated requires constant updates as IANA and RIRs keep allocating prefixes. Therefore

automation of such prefix filters is key for the success of this approach. One should probably not consider solutions described in this section if it is not capable of maintaining updated prefix filters: damage would probably be worse than the intended security policy.

5.1.2.1. 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 [24]. 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 [27]. 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.

If process in place (manual or automatic) cannot guarantee that the list is updated regularly then it's better not to configure any filter based on allocated networks. The IPv4 experience has shown that many network operators implemented filters for prefixes not allocated by IANA but did not update them on a regular basis. This created problems for latest allocations and required a extra work for RIR's that had to "de-boggonize" the newly allocated prefixes.

5.1.2.2. 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.

5.1.2.3. Prefix filters creation from Internet Routing Registries (IRR)

An Internet Routing Registry (IRR) is a database containing internet routing information, described using Routing Policy Specification Language objects [14]. Network engineers are given privileges to describe routing policies of their own networks in the IRR and information is published, usually publicly. Most of Regional Internet Registries do also operate an IRR and can control that registered routes conform to allocations made.

It is possible to use IRR information in order to build for a given BGP neighbor a list of prefixes, 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 IRR 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 IRR to poll. Some IRR have been created and are not restricted to a given region or authoritative RIR. They allow RIRs to publish information contained in their IRR in a common place. They also make it possible for any subscriber (probably under contract) to publish information too. When doing requests inside such an IRR, 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 [28] (Routing Assets Database).

As objects in IRR's 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.

5.1.2.4. 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.

5.1.3. Prefixes too specific

Most ISPs will not accept advertisements beyond a certain level of specificity (and in return do not announce prefixes they consider as too specific). That acceptable specificity is decided for each peering between the 2 BGP peers. Some ISP communities have tried to document acceptable specificity. This document does not make any judgement on what the best approach is, it just recalls that there are existing practices on the internet and recommends the reader to refer to what those are. As an example RIPE community has documented that IPv4 prefixes longer than /24 and IPv6 prefixes longer than /48 are generally not announced/accepted in the internet [21] [22].

5.1.4. Filtering prefixes belonging to local AS

A network SHOULD filter its own prefixes on peerings with all its peers (inbound direction). This prevents local traffic (from a local source to a local destination) to leak over an external peering in case someone else is announcing the prefix over the Internet. This also protects the infrastructure which may directly suffer in case

backbone's prefix is suddenly preferred over the Internet. To an extent, such filters can also be configured on a network for the prefixes of its downstreams in order to protect them too. 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.

5.1.5. Internet exchange point (IXP) LAN prefixes

5.1.5.1. Network security

When a network is present on an exchange point (IXP) and peers with other IXP members over a common subnet (IXP LAN prefix), it MUST NOT accept more specific prefixes for the IXP LAN prefix from any of all its external BGP peers. Accepting these routes would create a black hole for connectivity to the IXP LAN.

If the IXP LAN prefix is accepted as an "exact match", care needs to be taken to avoid other routers in the network sending IXP traffic towards the externally-learned IXP LAN prefix (recursive route lookup pointing into the wrong direction). This can be achieved by preferring IGP routes before eBGP, or by using "BGP next-hop-self" on all routes learned on that IXP.

If the IXP LAN prefix is accepted at all, it MUST only be accepted from the ASes that the IXP authorizes to announce it - which will usually be automatically achieved by filtering announcements by IRR DB.

5.1.5.2. pMTUd and loose uRPF problem

In order to have pMTUd working in the presence of loose uRPF, it is necessary that all the networks that may source traffic that could flow through the IXP (ie. IXP members and their downstreams) have a route for the IXP LAN prefix. This is necessary as "packet too big" ICMP messages sent by IXP members' routers may be sourced using an address of the IXP LAN prefix. In the presence of loose uRPF, this ICMP packet is dropped if there is no route for the IXP LAN prefix or a less specific route covering IXP LAN prefix.

In that case, any IXP member SHOULD make sure it has a route for the IXP LAN prefix or a less specific prefix on all its routers and that it announces the IXP LAN prefix or less specific (up to a default route) to its downstreams. The announcements done for this purpose

SHOULD pass IRR-generated filters described in Section 5.1.2.3 as well as "prefixes too specific" filters described in Section 5.1.3. The easiest way to implement this is that the IXP itself takes care of the origination of its prefix and advertises it to all IXP members through a BGP peering. Most likely the BGP route servers would be used for this. The IXP would most likely send its entire prefix which would be equal or less specific than the IXP LAN prefix.

5.1.5.3. Example

Let's take as an example an IXP in RIPE region for IPv4. It would be allocated a /22 by RIPE NCC (X.Y.0.0/22 in our example) and use a /23 of this /22 for the IXP LAN (let say X.Y.0.0/23). This IXP LAN prefix is the one used by IXP members to configure eBGP peerings. The IXP could also be allocated an AS number (AS64496 in our example).

Any IXP member MUST make sure it filters prefixes more specific than X.Y.0.0/23 from all its eBGP peers. If it received X.Y.0.0/24 or X.Y.1.0/24 this could seriously impact its routing.

The IXP SHOULD originate X.Y.0.0/22 and advertise it to its members through its BGP route servers (configured with AS64496).

The IXP members SHOULD accept the IXP prefix only if it passes the IRR generated filters (see Section 5.1.2.3)

IXP members SHOULD then advertise X.Y.0.0/22 prefix to their downstreams. This announce would pass IRR based filters as it is originated by the IXP.

5.1.6. Default route

5.1.6.1. IPv4

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

5.1.6.2. IPv6

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

5.2. 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 system 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.

5.2.1. Filters with internet peers

5.2.1.1. Inbound 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.

5.2.1.1.1. Inbound filtering loose option

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

- Prefixes not routable (Section 5.1.1)
- Prefixes not allocated by IANA (IPv6 only) (Section 5.1.2.1)
- Routes too specific (Section 5.1.3)
- Prefixes belonging to local AS (Section 5.1.4)
- Exchange points LAN prefixes (Section 5.1.5)
- Default route (Section 5.1.6)

5.2.1.1.2. Inbound filtering strict option

In that case, filters are applied to make sure advertisements strictly conform to what is declared in routing registries Section 5.1.2.2. 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 5.1.1)
- Routes too specific (Section 5.1.3)
- Prefixes belonging to local AS (Section 5.1.4)

- Exchange points LAN prefixes (Section 5.1.5)

- Default route (Section 5.1.6)

5.2.1.2. Outbound 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 5.1.1)

- Routes too specific (Section 5.1.3)

- Exchange points LAN prefixes (Section 5.1.5)

- Default route (Section 5.1.6)

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.

5.2.2. Filters with customers

5.2.2.1. Inbound filtering

Inbound policy with end customers is pretty straightforward: only customers prefixes must be accepted, all others MUST 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.

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 inbound/outbound prefix 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 5.2.1.1 can be applied.

5.2.2.2. Outbound filtering

Outbound 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 5.1.1)

- Routes too specific (Section 5.1.3)

- Default route (Section 5.1.6)

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.

5.2.3. Filters with upstream providers

5.2.3.1. Inbound 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 5.2.1.1. 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.

5.2.3.2. Outbound filtering

The filters to be applied should not differ from the ones applied for internet peers (Section 5.2.1.2).

5.3. Prefix filtering recommendations for leaf networks

5.3.1. Inbound 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 5.1.6). 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 5.1.1)

- Routes too specific (Section 5.1.3)

- Prefixes belonging to local AS (Section 5.1.4)

- Default route (Section 5.1.6) depending if the route is requested or not

5.3.2. Outbound 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 5.2.1.2 to avoid announcing invalid routes to its upstream provider.

6. 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 [20]. Author of this document proposes to follow the proposal of the RIPE community.

7. 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 plus some headroom to permit growth.

- 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 regularly review 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.

8. 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 is not adding anything extra in case prefix filters are built from registries as described in Section 5.1.2.3.

- 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 [30] (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.

9. Next-Hop Filtering

If peering on a shared network, like an Exchange-Point, BGP can advertise prefixes with a 3rd-party next-hop, thus directing packets not to the peer announcing the prefix but somewhere else.

This is a desirable property for BGP route-server setups [30], where the route-server will relay routing information, but has neither capacity nor desire to receive the actual data packets. So the BGP route-server will announce prefixes with a next-hop setting pointing to the router that originally announced the prefix to the route-server.

In direct peerings between ISPs, this is undesirable, as one of the peers could trick the other one to send packets into a black hole (unreachable next-hop) or to an unsuspecting 3rd party who would then have to carry the traffic. Especially for black-holing, the root cause of the problem is hard to see without inspecting BGP prefixes at the receiving router at the IXP.

Therefore, the authors recommend to, by default, apply an inbound route policy to IXP peerings which sets the next-hop for accepted prefixes to the BGP peer that sent the prefix (which is what "next-hop-self" would enforce on the sending side, but you can not rely on the other party to always send correct information).

This policy MUST NOT be used on route-server peerings, or on peerings where you intentionally permit the other side to send 3rd-party next-hops.

10. 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.

11. Change logs

11.1. Diffs between draft-jdurand-bgp-security-01 and draft-jdurand-bgp-security-00

Following changes have been made since previous document draft-jdurand-bgp-security-00:

- "This documents" typo corrected in the former abstract

- Add normative reference for RFC5082 in former section 3.2

- "Non routable" changed in title of former section 4.1.1

- Correction of typo for IPv4 loopback prefix in former section 4.1.1.1

- Added shared transition space 100.64.0.0/10 in former section 4.1.1.1

- Clarification that 2002::/16 6to4 prefix can cross network boundaries in former section 4.1.1.2

- Rationale of 2000::/3 explained in former section 4.1.1.2

- Added 3FFE::/16 prefix forgotten initially in the simplified list of prefixes that MUST not be routed by definition in former section 4.1.1.2

- Warn that filters for prefixes not allocated by IANA must only be done if regular refresh is guaranteed, with some words about the IPv4 experience, in former section 4.1.2.1

- Replace RIR database with IRR. A definition of IRR is added in former section 4.1.2.2

- Remove any reference to anti-spoofing in former section 4.1.4

- Clarification for IXP LAN prefix and pMTUd problem in former section 4.1.5

- "Autonomous filters" typo (instead of Autonomous systems) corrected in the former section 4.2

- Removal of an example for manual address validation in former section 4.2.2.1

- RFC5735 obsoletes RFC3300

- Ingress/Egress replaced by Inbound/Outbound in all the document

11.2. Diffs between draft-jdurand-bgp-security-02 and draft-jdurand-bgp-security-01

Following changes have been made since previous document draft-jdurand-bgp-security-01:

- 2 documentation prefixes were forgotten due to errata in RFC5735. But all prefixes were removed from that document which now point to other references for sake of not creating a new "registry" that would become outdated sooner or later.

- Change MD5 section with global TCP security session and introducing TCP-AO in former section 3.1. Added reference to BCP38

- Added new section 3 about BGP router protection with forwarding plane ACL

- Change text about prefix acceptable specificity in former section 4.1.3 to explain this doc does not try to make recommendations

- Refer as much as possible to existing registries to avoid creating a new one in former section 4.1.1.1 and 4.1.1.2

- Abstract reworded

- 6to4 exception described (only more specifics must be filtered)

- More specific -> more specifics

- should -> MUST for the prefixes an ISP needs to filter from its customers in former section 4.2.2.1

- Added "plus some headroom to permit growth" in former section 7

- Added new section on Next-Hop filtering

12. Acknowledgements

Authors would like to thank the following people for their comments and support: Marc Blanchet, Ron Bonica, Daniel Ginsburg, David Groves, Tim Kleefass, Hagen Paul Pfeifer, Thomas Pinaud, Carlos Pignataro, Matjaz Straus, Tony Tauber, Gunter Van de Velde, Sebastian

Wiesinger.

13. IANA Considerations

This memo includes no request to IANA.

14. Security Considerations

This document is entirely about BGP operational security.

15. References

15.1. 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] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

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

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

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

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

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

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

[9] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

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

15.2. Informative References

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

[12] 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.

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

[14] Blunk, L., Damas, J., Parent, F., and A. Robachevsky, "Routing Policy Specification Language next generation (RPSLng)", RFC 4012, March 2005.

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

[16] Blanchet, M., "Special-Use IPv6 Addresses", RFC 5156, April 2008.

[17] Cotton, M. and L. Vegoda, "Special Use IPv4 Addresses", BCP 153, RFC 5735, January 2010.

[18] Arkko, J., Cotton, M., and L. Vegoda, "IPv4 Address Blocks Reserved for Documentation", RFC 5737, January 2010.

[19] Dugal, D., Pignataro, C., and R. Dunn, "Protecting the Router Control Plane", RFC 6192, March 2011.

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

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

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

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

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

[25] "IANA IPv6 Address Space", <http://www.iana.org/assignments/

ipv6-address-space/ipv6-address-space.xml>.

[26] "IANA IPv6 Special Purpose Registry", <http://www.iana.org/assignments/iana-ipv6-special-registry/iana-ipv6-special-registry.xml>.

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

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

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

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

[31] "IANA Reserved IPv4 Prefix for Shared Address Space", <http://tools.ietf.org/id/draft-weil-shared-transition-space-request-15.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