

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 10, 2014

Camilo Cardona
Pierre Francois
IMDEA Networks
July 9, 2013

Making BGP filtering a habit: Impact on policies
draft-cardona-filtering-threats-02

Abstract

Network operators define their BGP policies based on the business relationships that they maintain with their peers. By limiting the propagation of BGP prefixes, an autonomous system avoids the existence of flows between BGP peers that do not provide any economical gain. This draft describes how undesired flows can emerge in autonomous systems due to the filtering of overlapping BGP prefixes by neighboring domains.

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 January 10, 2014.

Copyright Notice

Copyright (c) 2013 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 | 3 |
| 2. Filtering overlapping prefixes | 3 |
| 2.1. Local filtering | 4 |
| 2.2. Remotely triggered filtering | 5 |
| 3. Uses of overlapping prefix filtering that create undesired traffic flows | 6 |
| 3.1. Undesired Traffic Flows | 7 |
| 3.1.1. Undesired traffic flows caused by local filtering of overlapping prefixes | 8 |
| 3.1.2. Undesired traffic flows caused by remotely triggered filtering of overlapping prefixes | 11 |
| 4. Techniques to detect undesired traffic flows caused by filtering of overlapping prefixes | 14 |
| 4.1. Being the 'victim' | 15 |
| 4.2. Being a contributor to the existence of undesired traffic flows in other networks | 15 |
| 5. Techniques to counter undesired traffic flows due to the filtering of overlapping prefixes | 16 |
| 5.1. Reactive counter-measures | 17 |
| 5.2. Anticipant counter-measures | 18 |
| 5.2.1. Access lists | 18 |
| 5.2.2. Automatic filtering | 18 |
| 5.2.3. Neighbor-specific forwarding | 18 |
| 6. Conclusions | 19 |
| 7. References | 19 |
| Authors' Addresses | 19 |

1. Introduction

It is common practice for network operators to propagate overlapping prefixes along with the prefixes that they originate. It is also possible for some Autonomous Systems (ASes) to apply different policies to the overlapping (more specific) and the covering (less specific) prefix. Some ASes can even benefit from filtering the overlapping prefixes.

BGP makes independent, policy driven decisions for the selection of the best path to be used for a given IP prefix. However, routers must forward packets using the longest-prefix-match rule, which "precedes" any BGP policy (RFC1812 [4]). Indeed, the existence of a prefix p that is more specific than a prefix p' in the Forwarding Information Base (FIB) will let packets whose destination matches p be forwarded according to the next hop selected as best for p (the overlapping prefix). This process takes place by disregarding the policies applied in the control plane for the selection of the best next-hop for p' (the covering prefix). When overlapping prefixes are filtered and packets are forwarded according to the covering prefix, the discrepancy in the routing policies applied to covering and overlapping prefixes can create undesired traffic flows that infringe the policies of Internet Service Providing (ISPs) still holding a path towards the overlapping prefix.

This document presents examples of such cases and discusses solutions to the problem. The objective of this draft is to shed light on the use of prefix filtering by making the routing community aware of the cases where the effects of filtering might turn to be negative for the business of ISPs.

The rest of the document is organized as follows: Section 2 illustrates the motivation to filter overlapping prefixes. In Section 3, we provide some scenarios in which the filtering of overlapping prefixes lead to the creation of undesired traffic flows on other ASes. Section 4 and Section 5 discuss some techniques that ASes can use for, respectively, detect and react to undesired traffic flows.

2. Filtering overlapping prefixes

There are several scenarios where filtering an overlapping prefix is relevant to the operations of an AS. In this section, we provide examples of these scenarios. We differentiate cases in which the filtering is performed locally from those where the filtering is triggered remotely. These scenarios will be used as a base in Section 3 for describing side effects bound with such practices.

2.1. Local filtering

Let us first analyze the scenario depicted in Figure 1. AS1 and AS2 are two autonomous systems spanning a large geographical area and peering in 3 different physical locations. Let AS1 announce prefix 10.0.0.0/22 over all peering links with AS2. Additionally, let us define that there is part of AS1's network which exclusively uses prefix 10.0.0.0/24 and which is closer to a peering point than to others.

To receive the traffic destined to prefix 10.0.0.0/24 on the link closer to this subnet, AS1 could announce the overlapping prefix only over this specific session. At the time of the establishment of the peering, it can be defined by both ASes that hot potato routing would happen in both directions of traffic. In other words, it was agreed that each AS will deliver the traffic to the other AS on the nearest peering link. In this scenario, it becomes relevant to AS2 to enforce such practice by detecting the described situations and automatically issuing the appropriate filtering. In this case, by implementing these automatic procedures, AS2 would legitimately detect and filter prefix 10.0.0.0/24.

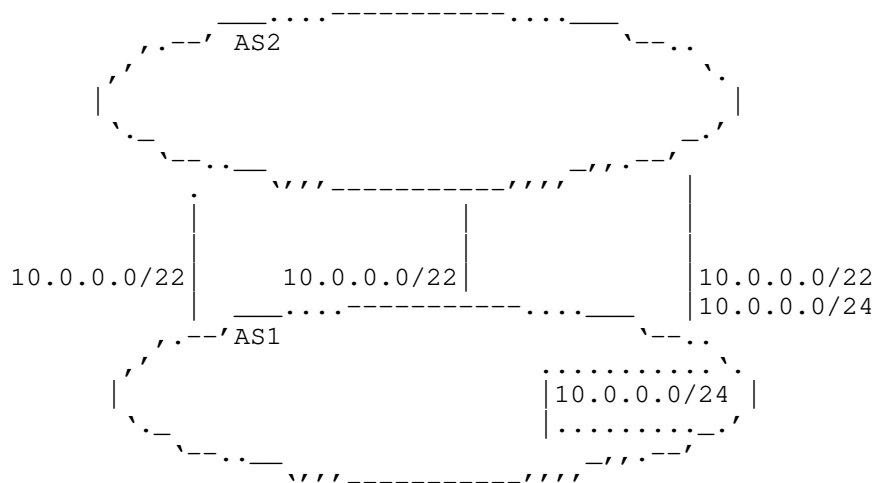


Figure 1: Basic scenario of local filtering

Local filtering could be required in other cases. For example, a dual homed AS receiving an overlapping prefix from only one of its providers. Figure 2 depicts a simple example of this case.

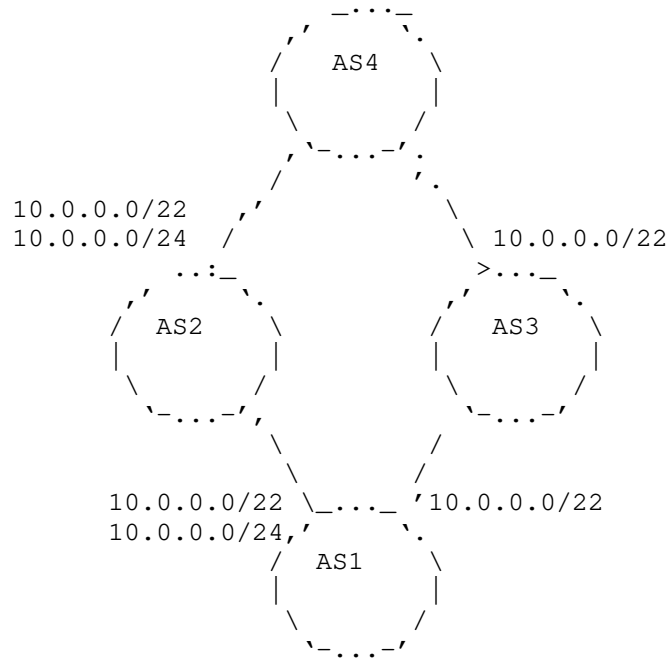


Figure 2: Basic scenario of local filtering

In this scenario, prefix 10.0.0.0/22 is advertised by AS1 to AS2 and AS3. Both ASes propagate the prefix to AS4. Additionally, AS1 advertises prefix 10.0.0.0/24 to AS2, which subsequently propagates the prefix to AS4.

It is possible that AS4 resolves to filter the more specific prefix 10.0.0.0/24. One potential motivation could be the economical preference of the path via AS2 over AS3. Another feasible reason is the existence of a technical policy by AS4 of aggregating incoming prefixes longer than /23.

The above examples illustrate two of the many motivations to configure routing within an AS with the aim of ignoring more specific prefixes. Operators have reported applying these filters in a manual fashion [3]. The relevance of such practice led to investigate automated filtering procedures in I-D.WHITE [5].

2.2. Remotely triggered filtering

ISPs can tag the BGP paths that they propagate to neighboring ASes with communities, in order to tweak the propagation behavior of the

ASes that handle these paths [1].

Some ISPs allow their direct and indirect customers to use such communities to let the receiving AS not export the path to some selected neighboring AS. By combining communities, the prefix could be advertised only to a given peer of the AS providing this feature. Figure 3 illustrates an example of this case.

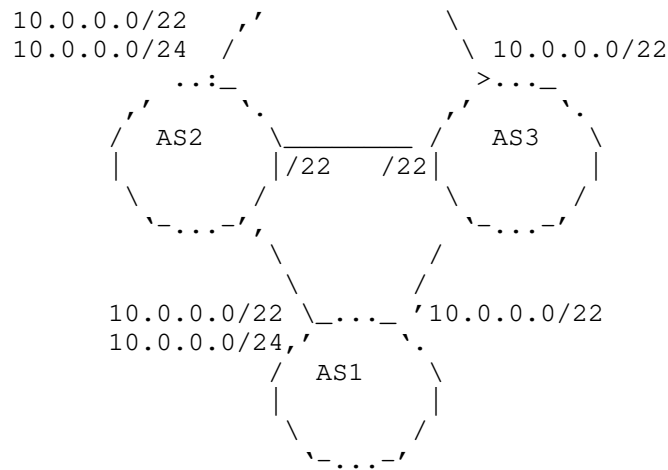


Figure 3: Remote triggered filtering

AS2 and AS3 are peers. Both ASes are providers of AS1. For traffic engineering purposes, AS1 could use communities to prevent AS2 from announcing prefix 10.0.0.0/24 to AS3.

Such technique is useful for operators to tweak routing decisions in order to align with complex transit policies. We will see in later sections that by producing the same effect as filtering, they can also lead to undesired traffic flows at other, distant, ASes.

3. Uses of overlapping prefix filtering that create undesired traffic flows

In this section we define the concept of undesired traffic flows and describe three configuration scenarios that lead to their creation. Note that these examples do not capture all the cases where such issues can take place. More examples will be provided in future revisions of this document.

3.1. Undesired Traffic Flows

The BGP policy of an Internet Service provider includes all actions performed over its originated routes and the routes received externally. One important part of the BGP policy is the selection of the routes that are propagated to each neighboring AS. One of the goals of these policies is to allow ISPs to avoid transporting traffic between two ASes without economical gain. For instance, ISPs typically propagate to their peers only routes coming from its customers (RFC4384 [6]). We briefly illustrate this operation in Figure 4. In the figure, AS2 is establishing a settlement free peering with AS1 and AS3. AS2 receives prefix P3/p3, from AS3. AS2, however, is not interested in transporting traffic from AS1 to AS3, therefore it does not propagate the prefix to AS3. In the figure, we also show a customer of AS2, AS4, which is announcing prefix P4/p4. AS2 propagates this prefix to AS1.

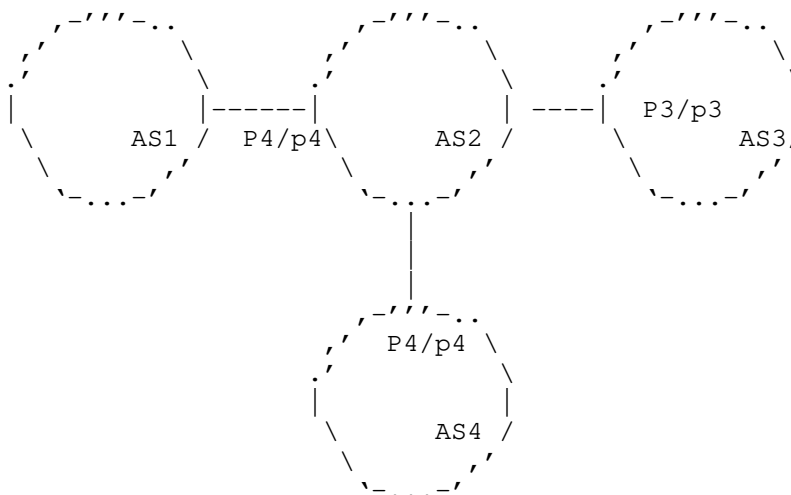


Figure 4: Prefix exchange among four autonomous systems

Although ISPs usually implement the aforementioned policies, undesired traffic flows may still appear. In Figure 4, undesired traffic flows are created, when, despite AS2's policy, traffic arriving from peer AS1 is received and transported to AS3 by AS2. These type of traffic flows can arise due to a number of reasons. Specifically, in this document we explain how the filtering of overlapping prefixes might cause undesired traffic flows on ASes. We provide examples of these cases in the next sections.

3.1.1. Undesired traffic flows caused by local filtering of overlapping prefixes

In this section we describe cases in which an AS locally filters an overlapping prefix. We show that, depending on the BGP policies applied by surrounding ASes, this decision can lead to undesired traffic flows.

3.1.1.1. Initial setup

We start by describing the basic scenario of this case in Figure 5.

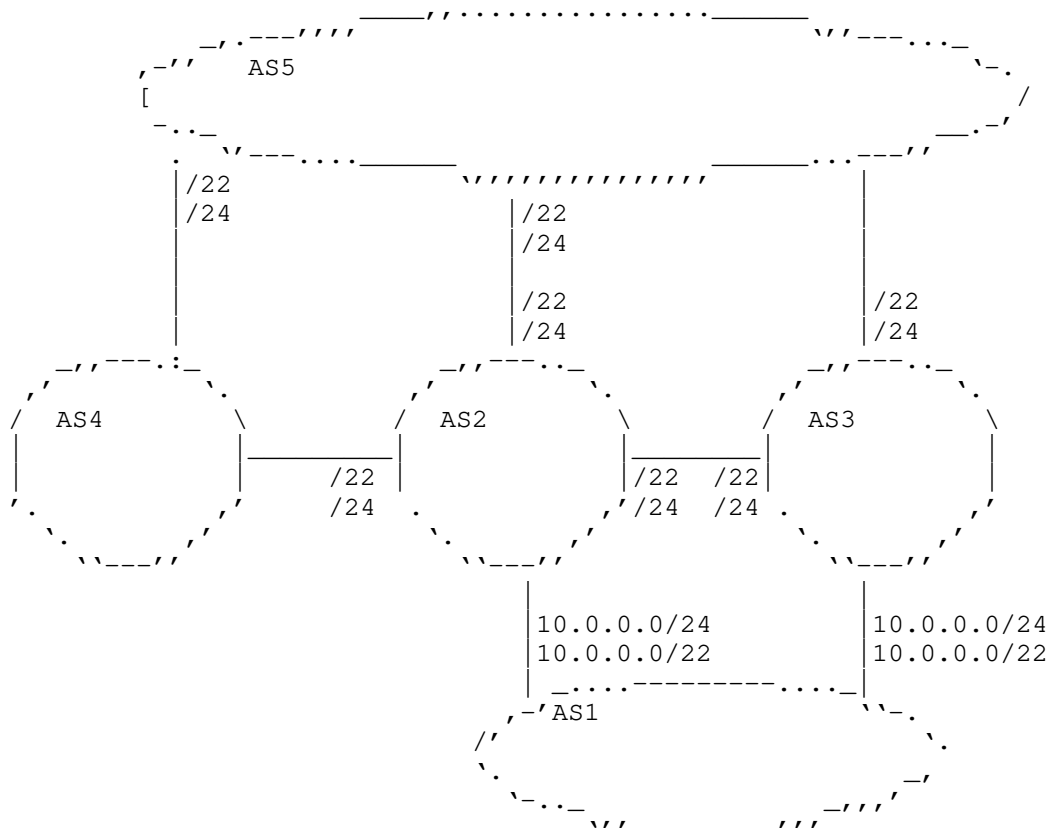


Figure 5: Initial Setup Local

AS1 is a customer of AS2 and AS3. AS2, AS3, and AS4 are customers of AS5. AS2 is establishing a peering with AS3 and AS4. AS1 is announcing a covering prefix, 10.0.0.0/22, and an overlapping prefix

10.0.0.0/24 to its providers. In the initial setup, AS2 and AS3 announce the two prefixes to their peers and transit providers. AS4 receives both prefixes from its peer (AS2) and transit provider (AS5). We will consider that AS5 chooses the path through AS3 to reach AS1.

3.1.1.2. Undesired traffic flows by local filtering - Case 1

In the next scenarios, we show that if AS4 filters the incoming overlapping prefix from AS5, there is a situation in which undesired traffic flows are created on other ASes.

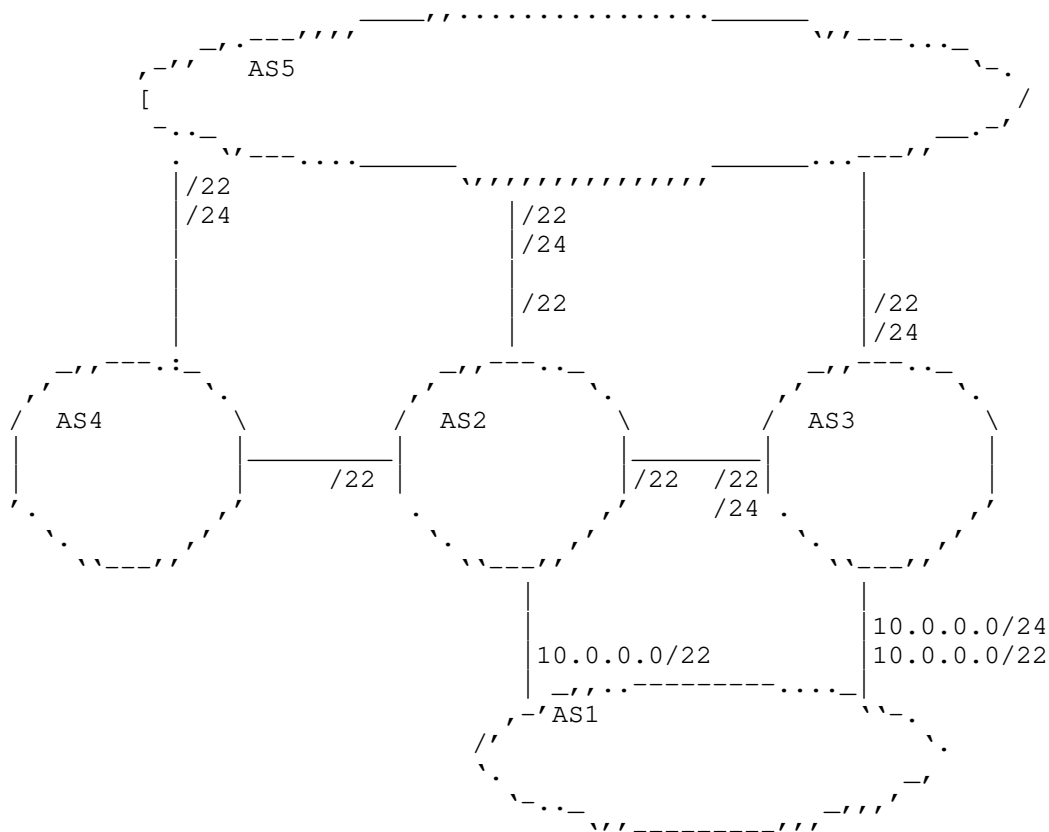


Figure 6: Undesired traffic flows by local filtering - Case 1

Let us assume the scenario illustrated in Figure 6. For this case, AS1 only propagates the overlapping prefix to AS3. AS4 receives the overlapping prefix only from its transit provider, AS5.

AS4 now is in a situation in which it would be favorable for it to filter the announcement of prefix 10.0.0.0/24 from AS5. Subsequently, traffic from AS4 to prefix 10.0.0.0/24 is forwarded towards AS2. Because AS2 receives the more specific prefix from AS3, traffic from AS4 to prefix 10.0.0.0/24 follows the path AS4-AS2-AS3-AS1. AS2's BGP policies are implemented to avoid AS2 to exchange traffic between AS4 and AS3. However, due to the discrepancies of routes from the overlapping and covering prefixes, undesired traffic flows between AS4 and AS3 still exist on AS2's network. This situation is economically detrimental for AS2, since it forwards traffic from a peer to a non-customer neighbor.

3.1.1.3. Undesired traffic flows by local filtering - Case 2

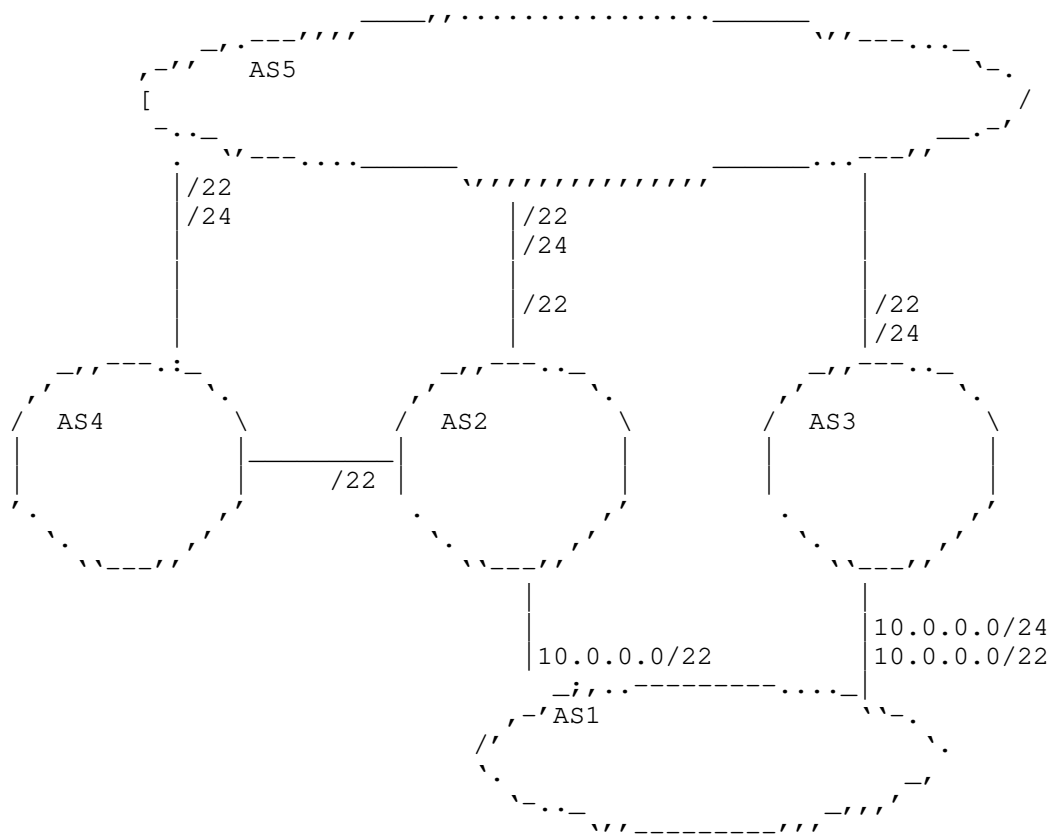


Figure 7: Undesired traffic flows after local filtering - Case 2

Let us assume a second case where AS2 and AS3 are not peering and AS1 only propagates the overlapping prefix to AS3. AS4 receives the overlapping prefix only from its transit provider, AS5. This case is illustrated in Figure 7.

Similar to the scenario described in Section 3.1.1.2, AS4 is in a situation in which it would be favorable to filter the announcement of prefix 10.0.0.0/24 from AS5. Subsequently, traffic from AS4 to prefix 10.0.0.0/24 is forwarded towards AS2. Due to the existence of a route to prefix 10.0.0.0/24, AS2 receives the traffic heading to this prefix from AS4, and sends it to AS5. This situation creates undesired traffic flows that contradict AS2's BGP policy, since the AS ends up forwarding traffic from a peer to a transit network.

3.1.2. Undesired traffic flows caused by remotely triggered filtering of overlapping prefixes

We present a configuration scenario in which an AS, using the mechanism described in Section 2.2, informs its provider to selectively propagate an overlapping prefix, leading to the creation of undesired traffic flows in another AS.

3.1.2.1. Initial setup

Let AS1 be a customer of AS2 and AS3. AS1 owns 10.0.0.0/22, which it advertises through AS2 and AS3. Additionally, AS2 and AS3 are peers.

Both AS2 and AS3 select AS1's path as best, and propagate it to their customers, providers, and peers. Some remote ASes will route traffic destined to 10.0.0.1 through AS2 while others will route traffic through AS3.

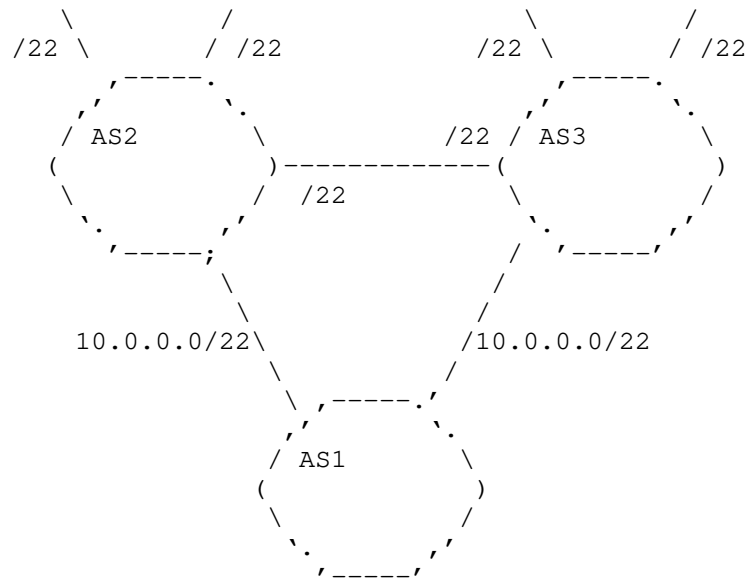


Figure 8: Example scenario

3.1.2.2. Injection of an overlapping prefix

Let AS1 advertise 10.0.0.0/24 over AS3 only. AS3 would propagate this prefix to its customers, providers, and peers, including AS2.

From AS2's point of view, the path towards 10.0.0.0/24 is a "peer path" and AS2 will only advertise it to its customers. ASes in the customer branch of AS2 will receive a path to the /24 that contains AS3 and AS2. Some multi-homed customers of AS2 may also receive a path through AS3, but not through AS2, from other peering or provider links. Any remote AS that is not lying in the customer branch of AS2, will receive a path for 10.0.0.0/24 through AS3 and not through AS2.

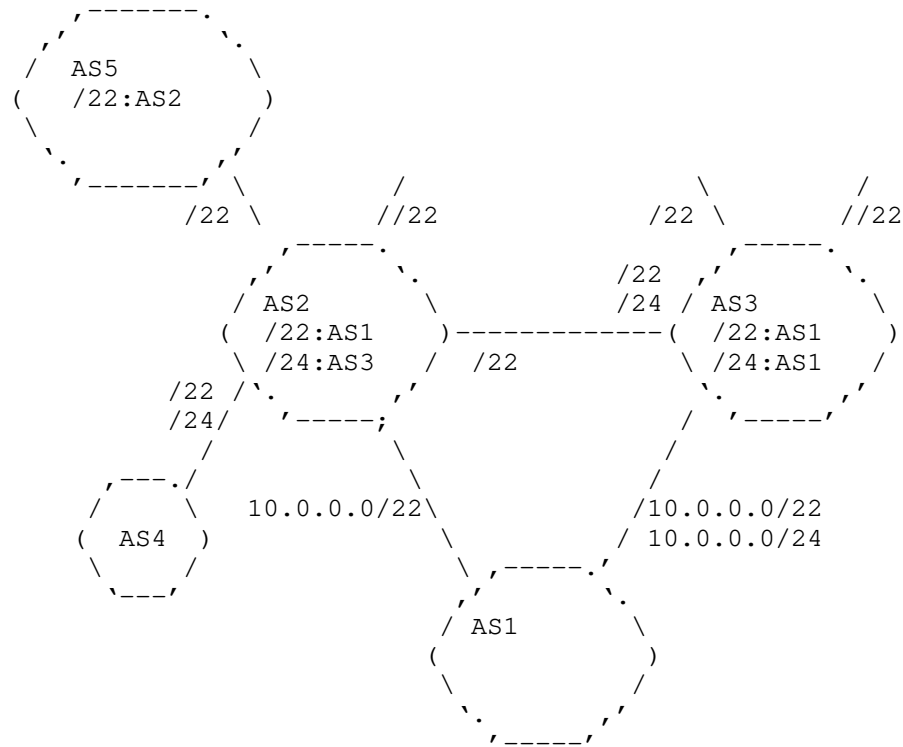


Figure 10: More Specific Injection

From AS2's point of view, such a path is a "peer path" and will only be advertised by AS2 to its customers.

ASes that are not customers of AS2 will not receive a path for 10.0.0.0/24. These ASes will forward packets destined to 10.0.0.0/24 according to their routing state for 10.0.0.0/22. Let us assume that AS5 is such an AS, and that its best path towards 10.0.0.0/22 is through AS2. Then, packets sent towards 10.0.0.1 by AS5 will eventually reach AS2. However, in the data-plane of the nodes of AS2, the longest prefix match for 10.0.0.1 is 10.0.0.0/24, which is reached through AS3, a peer of AS2. Since AS5 is not in the customer branch of AS2, we are in a situation in which traffic flows between non-customer AS take place in AS2.

4. Techniques to detect undesired traffic flows caused by filtering of overlapping prefixes

We differentiate the techniques available for detecting undesired

traffic flows caused by the described scenarios from the cases in which the interested AS is the victim or contributor of such operations.

4.1. Being the 'victim'

To detect if undesired traffic flows are taking place in its network, an ISP can monitor its traffic data and validate if any flow entering the ISP network through a non-customer link is forwarded to a non-customer next-hop.

As mentioned in Section 3.1, undesired traffic flows might appear due to different situations. To discover if the problem arose after the filtering of prefixes by neighboring ASes, an operator can analyze available BGP data. For instance, an ISP can seek for overlapping prefixes for which the next-hop is through a provider (or peer), while the next-hop for their covering prefix(es) is through a client. Direct communication or looking glasses can be used to check whether non-customer neighboring ASes are propagating a path towards the covering prefix to their own customers, peers, or providers. This should trigger a warning, as this would mean that ASes in the surrounding area of the current AS are forwarding packets based on the routing entry for the less specific prefix only.

4.2. Being a contributor to the existence of undesired traffic flows in other networks

It can be considered problematic to be causing undesired traffic flows on other ASes. This situation may appear as an abuse to the network resources of other ISPs.

There may be justifiable reasons for one ISP to perform filtering, either to enforce established policies or to provide prefix advertisement scoping features to its customers. These can vary from trouble-shooting purposes to business relationships implementations. Restricting such features for the sake of avoiding the creation of undesired traffic flows is not a practical option.

Traffic data does not help an ISP detect that it is acting as a contributor of the creation of the undesired traffic flow. It is thus advisable to obtain as much information as possible about the Internet environment of the AS and assess the risks of filtering overlapping prefixes before implementing them.

Monitoring the manipulation of the communities that implement the scoping of prefixes is recommended to the ISPs that provide these features. The monitored behavior should then be faced against their terms of use.

5. Techniques to counter undesired traffic flows due to the filtering of overlapping prefixes

Network Operators can adopt different approaches with respect to undesired traffic flows. We classify these actions according to whether they are anticipant or reactive.

Reactive approaches are those in which the operator tries to detect the situations and solve undesired traffic flows, manually, on a case-by-case basis.

Anticipant or preventive approaches are those in which the routing system will not let the undesired traffic flows actually take place when the configuration scenario is set up.

We will describe these two kinds of approaches in the following part of this Section. We will use the scenario depicted in Figure 11 to provide examples for the different techniques.

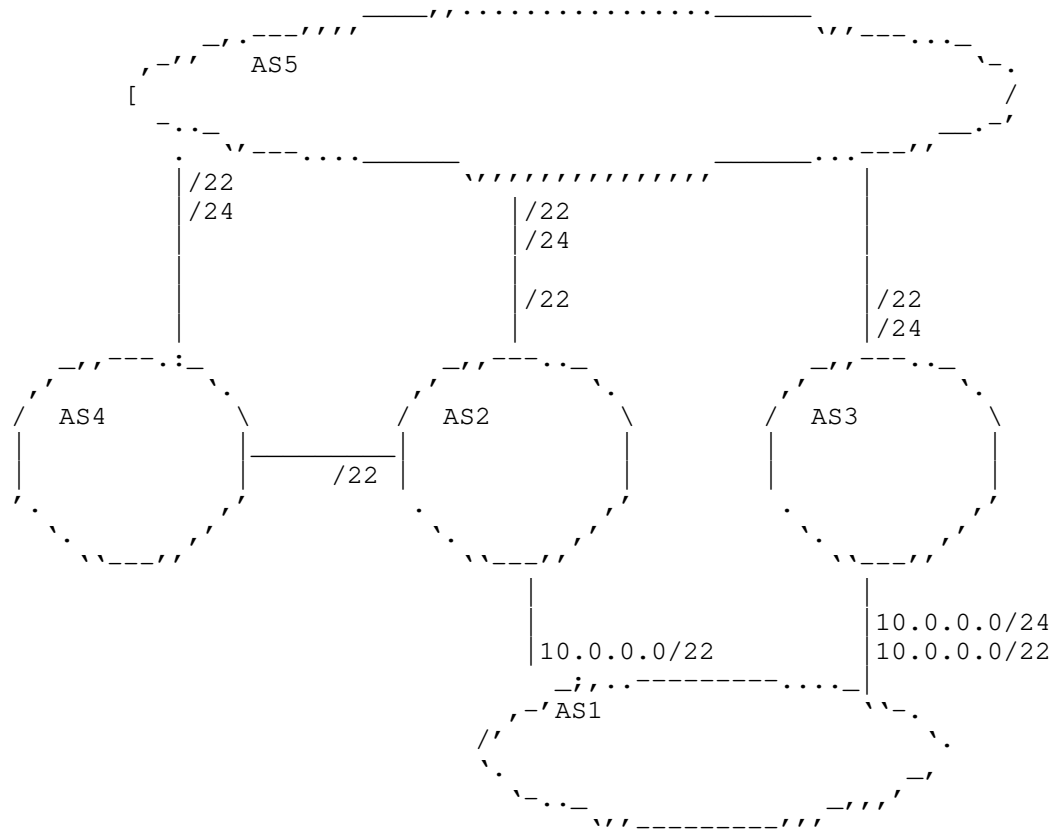


Figure 11: Anticipant counter-measures - Base example

5.1. Reactive counter-measures

An operator who detects that its policies are threatened by undesired traffic flows can contact the ASes that are likely to have performed the propagation tweaks, inform them of the situation and persuade them to change their behavior.

For some cases, if the external ASes maintain their behavior, an operator can account the amount of traffic that has been subject to the undesired flows and charge the peer for that traffic. That is, the operator can claim that it has been a provider of that peer for the traffic that transited between the two ASes.

An operator can decide to filter-out the concerned overlapping prefix at the peering session over which it was received. In the example of Figure 11, AS2 would filter out the incoming prefix 10.0.0.0/24 from

the eBGP session with AS5. As a result, the traffic destined to that /24 would be forwarded by AS2 along its link with AS1, despite the actions performed by AS1 to have this traffic coming in through its link with AS3.

5.2. Anticipant counter-measures

5.2.1. Access lists

An operator can configure its routers to install dynamically an access-list made of the prefixes towards which the forwarding of traffic from that interface would lead to undesired traffic flows. Note that this technique actually lets packets destined to a valid prefix be dropped while they are sent from a neighboring AS that cannot know about policy conflicts and hence had no means to avoid the creation of undesired traffic flows.

In the example of Figure 11, AS2 would install an access-list denying packets matching 10.0.0.0/24 associated with the interface connecting to AS4. As a result, traffic destined to that prefix would be dropped, despite the existence of a valid route towards 10.0.0.0/22.

5.2.2. Automatic filtering

As described in Section 3, filtering of overlapping prefixes can in some scenarios lead to undesired traffic flows. Nevertheless, depending on the autonomous system implementing such practice, this operation can prevent these cases. This can be illustrated using the example described in Figure 11: if AS2 or AS3 filter prefix 10.0.0.0/24, there would be no undesired traffic flow in AS2.

5.2.3. Neighbor-specific forwarding

An operator can technically ensure that traffic destined to a given prefix will be forwarded from an entry point of the network based only on the set of paths that have been advertised over that entry point.

As an example, let us analyze the scenario of Figure 11 from the point of view of AS2. The edge router connecting to the AS4 forward packets destined to prefix 10.0.0.0/24 towards AS5. Likewise, it will forward packets destined to prefix 10.0.0.0/22 towards AS1. The router, however, only propagates the path of the covering prefix (10.0.0.0/22) to AS4. An operator could implement the necessary techniques to force the edge router to forward packets coming from AS4 based only on the paths propagated to AS4. Thus, the edge router would forward packets destined to 10.0.0.0/24 towards AS1 in which case no undesired traffic flow would occur. This functionality could

be implemented in different ways. [2] describes an approach to implement this Behavior.

6. Conclusions

In this document, we described threats to policies of autonomous systems caused by the filtering of overlapping prefixes by external networks. We provide examples of scenarios in which undesired traffic flows are caused by these practices and introduce some techniques for their detection and prevention. We observe that there are reasonable situations in which ASes could filter overlapping prefixes, however, we encourage that network operators implement this type of filters only after considering the cases described in this document.

7. References

- [1] Donnet, B. and O. Bonaventure, "On BGP Communities", ACM SIGCOMM Computer Communication Review vol. 38, no. 2, pp. 55-59, April 2008.
- [2] Vanbever, L., Francois, P., Bonaventure, O., and J. Rexford, "Customized BGP Route Selection Using BGP/MPLS VPNs", Cisco Systems, Routing Symposium <http://www.cs.princeton.edu/~jrex/talks/cisconag09.pdf>, October 2009.
- [3] "INIT7-RIPE63", <<http://ripe63.ripe.net/presentations/48-How-more-specifics-increase-your-transit-bill-v0.2.pdf>>.
- [4] <<http://www.ietf.org/rfc/rfc1812.txt>>
- [5] <<http://tools.ietf.org/html/draft-white-grow-overlapping-routes-00>>
- [6] <<http://www.ietf.org/rfc/rfc4384.txt>>

Authors' Addresses

Camilo Cardona
IMDEA Networks
Avenida del Mar Mediterraneo, 22
Leganes 28919
Spain

Email: juancamilo.cardona@imdea.org

Pierre Francois
IMDEA Networks
Avenida del Mar Mediterraneo, 22
Leganes 28919
Spain

Email: pierre.francois@imdea.org

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 12, 2014

M. Chen
S. Zhuang
Huawei Technologies
Y. Zhu
S. Wang
China Telecom Co.,Ltd
July 11, 2013

Use Cases of Route Reflection based Traffic Steering
draft-chen-idr-rr-based-traffic-steering-usecase-00

Abstract

Route Reflection based Traffic Steering (RRTS) is an idea that leverages the BGP route reflection mechanism to realize traffic steering in the network, therefore the operators can conduct their traffic to transmit/receive through specific nodes, domains and/or planes as demand. This document introduces the requirements and use cases of RRTS.

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].

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 January 12, 2014.

Copyright Notice

Copyright (c) 2013 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. Problem Statement | 2 |
| 2. Use Cases and Requirements | 4 |
| 2.1. Multihoming Scenario | 4 |
| 2.2. Multiple Planes Scenario | 6 |
| 2.3. Multiple Entries/Exits Scenario | 7 |
| 2.4. Requirement Summary | 8 |
| 3. Route Refelection based Traffic Steering | 8 |
| 4. IANA Considerations | 10 |
| 5. Acknowledgements | 10 |
| 6. References | 10 |
| 6.1. Normative References | 10 |
| 6.2. Informative References | 10 |
| Authors' Addresses | 10 |

1. Problem Statement

In an IP network, typically, both the Interior Gateway Protocol (IGP) and Border Gateway Protocol (BGP) are simultaneously deployed to forward traffic from one domain to other domains. The IGP is responsible for the internal routing and connectivity, the BGP is responsible for inter-domain routing. For the inter-domain traffic, it is forwarded based on the BGP routes. But when the traffic enters a specific domain, since the BGP routes depend the IGP routes to reach to the BGP nexthop router, the traffic actually follows the IGP routes to reach to the Autonomous System Border Routers (ASBRs) and then is forwarded to next domain. So, the IGP topology, link metric and related policies determine the traffic path within the domain. Setting and adjusting the IGP metrics is the major practice method to conduct the traffic. In order to fully use the network bandwidth, reduce the congestion on links and/or nodes, the operators have to carefully design and adjust the IGP metrics. Design IGP metrics for a greenfield network is relatively easy. But for a product network, with the increasing of network size, density and traffic volume, it's hard or even impossible to adjust the IGP metrics to smoothly conduct the traffic as needed. Setting or changing IGP metric just likes a teeterboard, it often happens that changing the metric of one link to

solve one problem, and there will be another problem occurs. And even worse, bad metric design may result in route oscillation.

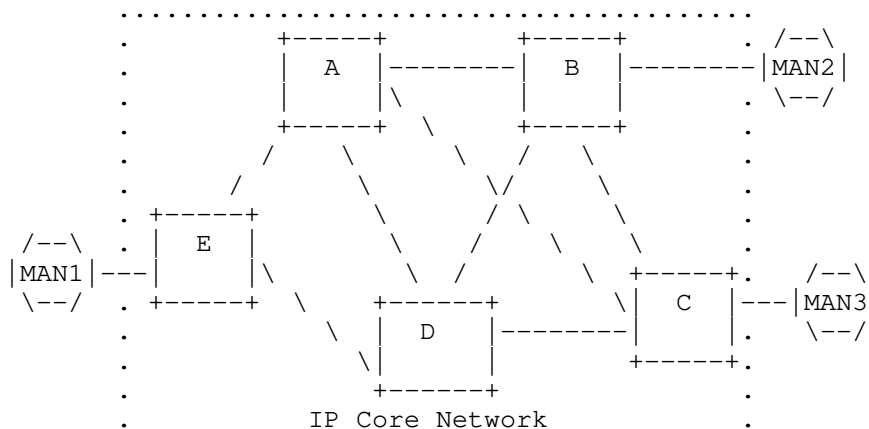


Figure 1: Paradoxical Metric

Figure 1 shows a paradoxical metric scenario. Where router A, B, C, D and E belong to the IP core network, Router A, B, C and D connect each other with full mesh links, and each link have the same metric; router E multihoming connects to router A and D. The Metropolitan Area Network 1 (MAN1) connects to the IP core network through router E, MAN2 connects to the IP core network through router B, and MAN3 connects to the IP core network through router C. The requirement would like this: the traffic between MAN1 and MAN2 is required to follow the path: E-A-B; and the traffic between MAN1 and MAN3 is required to follow the path: E-D-C. To satisfy the former requirement, it requires that the metric of link E-A must be less than the metric of link E-D. But to satisfy the later requirement, it will require that the metric of link E-A must be larger than the metric of link E-D. It's impossible to satisfy the paradoxical metric requirements simultaneously.

In addition, the existing BGP route decision is mainly based on the destination address, it does not consider the source address. From the source node point of view, the selected best route may not be the best route for the source node, especially in the network where Route Reflection is largely deployed. There are some proposed mechanisms (e.g., add-path[I-D.ietf-idr-add-paths], optimal route reflection [I-D.ietf-idr-bgp-optimal-route-reflection]) that may solve or mitigate the issues. But they also bring some new challenges, they will require more memory to save huge extra routes, to keep more states and make the implementation more complicated. The most

important one is that these solutions require to upgrade not just only one or two deployed devices, they may require to upgrade the whole or most the network devices. This makes it difficult to be deployed in a product network.

Route Reflection based Traffic Steering (RRTS) is an idea that leverages the BGP route reflection mechanism to realize traffic steering in the network, therefore the operators can conduct their traffic to transmit/receive through specific nodes, domains and/or planes as demand. The essential of RRTS is that the concept of traffic engineering is introduced into BGP network.

This document introduces some use cases and requirements of the RRTS.

2. Use Cases and Requirements

2.1. Multihoming Scenario

Figure 2 is a multihoming scenario, where the MANs are connected by an IP core network. The routers in the core network run both IGP (ISIS or OSPF) and BGP, the IGP is used to achieve the internal routing and connectivity. There will be full mesh I-BGP sessions among the core routers or I-BGP sessions between the routers and the Route Reflector (RR). There will be E-BGP sessions between the core routers of the IP core network and the edge routers of the MANs. The BGP is used to distribute the Internet and the MAN routes.

For each MAN, it multihoming connects to the IP core network through two or more core routers. Traffic between the MANs are typically forwarded through the IP core network. At the same time, there are some MANs that may have direct connected links between them.

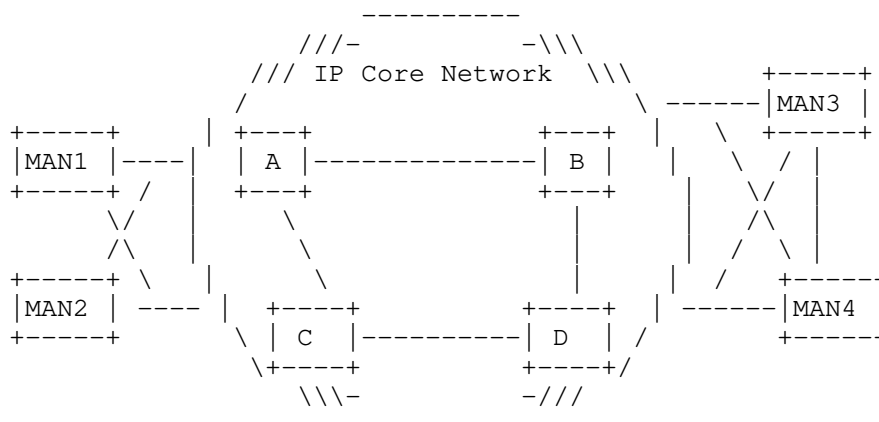


Figure 2: Multihoming Scenarios

For such a network, there will be requirements like this:

If there are direct links between MANs, the traffic should be forwarded through the direct connected links; and if the direct connected links are used up, then some traffic should be forwarded through the IP core network;

For two specific MANs (e.g., MAN1 and MAN3), the traffic between them should be forwarded through the required path, for example, MAN1-A-B-MAN3; the working and backup path should be disjoint as soon as possible.

For two specific MANs (e.g., MAN2 and MAN4), part of the traffic is required to be forwarded through one path (e.g., MAN2-C-D-MAN4), and other traffic is required to be forwarded through other path (e.g., MAN2-A-B-MAN4);

There may be more other requirements with the increasing of the network size, density, and more services transmitted over the network, more access networks connect to the network.

As discussed in the previous section, the current metric-based traffic conducting mechanism cannot (at least does not easily) satisfy the above requirements.

2.2. Multiple Planes Scenario

With the increasing of network traffic, the bandwidth, device ports of the existing devices/network are not enough to support new accessed traffic and services. So, some operators choose to set up new parallel planes to enlarge the network capacity.

Figure 3 is a multiple planes scenario, there are two planes in the IP core network. C11, C12, C13 and C14 belong to plane 1; C21, C22, C23 and C24 belong to plane 2. Plane 2 is the new built plane that normally has more bandwidth and is fine designed. So, the operators will move the high-cost services to the new plane, and keep the other services on the old plane, and try to fully use the network resource of the two planes and keep the traffic balanced according to the capacity of two planes.

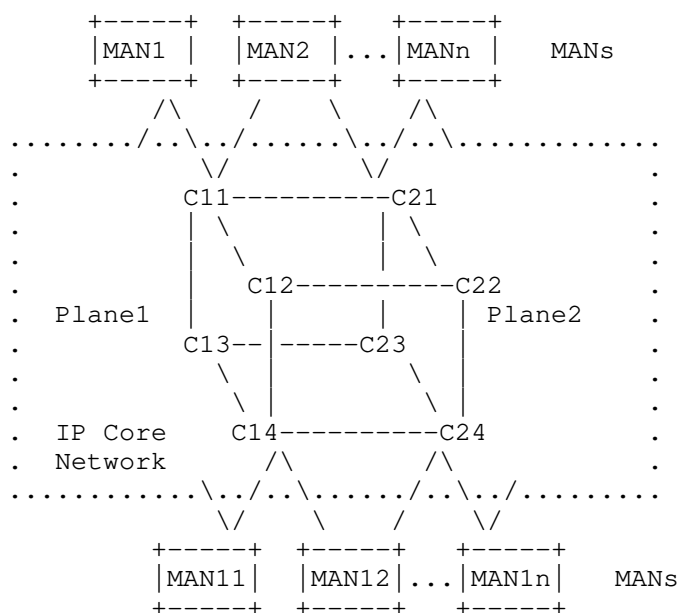


Figure 3: Multiple Planes Scenarios

For the multiple planes network, here are some typical requirements:

For any two specific MANs (e.g., MAN1 and MAN11), some service (e.g., Internet Data Center (IDC), Virtual Private Network (VPN), Private Line etc. services) traffic is required to be forwarded through the new plane (plane 2), and the other traffic will still be forwarded through the old plane (plane 1);

Traffic between some MANs is required to be forwarded through plane 1, and traffic between other MANs is required to be forwarded through plane 2; for example, traffic between MAN1 and MAN11 is required to be forwarded through plane 1, traffic between MAN3 and MAN33 is required to be forwarded through plane 2.

For any two specific MANs (e.g., MAN2 and MAN22), it should be able to balance the traffic between the two MANs through the two planes based on the capacity/load of the two planes;

According to different users, it should be able to choose different planes;

According different SLA and QoS requirements, it should be able to choose proper forwarding plane based on the SLA and QoS requirements and the fact of the planes;

It should be able to choose forwarding plane based on the different access locations;

2.3. Multiple Entries/Exits Scenario

Figure 4 shows the multiple entries/exits scenario. For network 1, it has three entries/exits that respectively connect to transit network A, B and C. And between network 1 and each transit network, there is one or more links. Different link has different cost/price, bandwidth, delay/loss attributes.

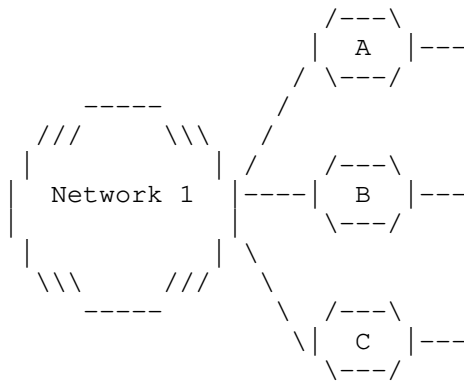


Figure 4: Multiple Entry/Exit (MEE)

For this multiple entry/exit scenario, it has the following requirements:

Choose the proper entry/exit based on link price and/or service type;

Dynamically adjust the entry/exit based on link load and/or link price;

2.4. Requirement Summary

According to the above use cases, the requirements can be summarized as follows:

Be able to specify the forwarding path/plane based source and destination addresses;

Be able to specify the forwarding path/plane based on service type;

Be able to specify the forwarding path/plane based on users;

Be able to specify the forwarding path/plane based on SLA and QoS requirements;

Be able to change/adjust the forwarding path/plane of some traffic based on the network load and usable capacity;

Be able to choose/adjust network entry/exit based on link price/service type/link load;

Looking through these requirements, they are actually the requirements of traffic engineering. In tradition IP network, traffic forwarding is a per-hop IP lookup and forwarding behavior. There is few mechanism defined for pure IP based traffic engineering. IP source routing is a way that can direct the traffic to transmit along specified path, but it is not widely implemented and deployed. That means, there is requirement to introduce the traffic engineering to pure IP network, but it is lack of readily available solutions.

3. Route Refelection based Traffic Steering

For a product network, an acceptable solution should be able to smoothly and incrementally upgrade the network and should not affect the on-going services. Route Reflection is widely deployed in the field, a Route Reflector (RR) has the ability to "install"/distribute a route to its client with the nexthop that can be either the RR itself or any other different BGP speakers. Given this, for an IP network, if all routers run BGP and are connected by a centralized RR, and the RR has the topology, network capacity, network resource etc. information of the whole network. Then the RR can compute the

routes for every router and install/distribute the routes to corresponding routers.

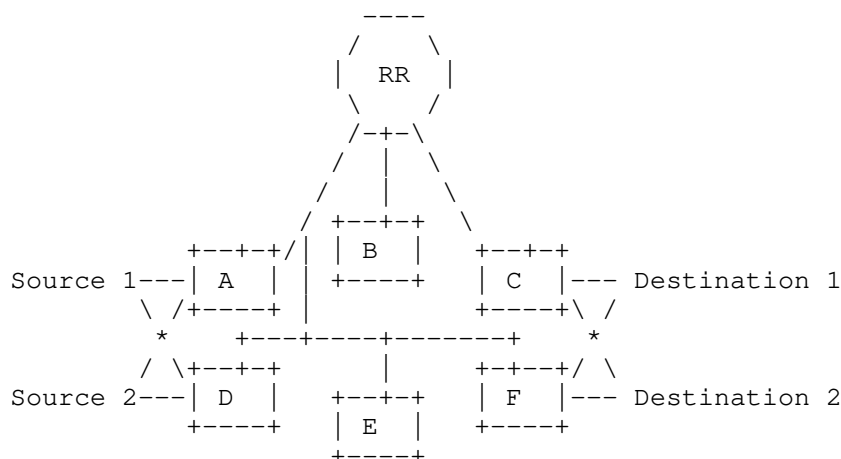


Figure 5: Route Reflection based Traffic Steering (RRTS)

Figure 5 is a reference architecture of the Route Reflection based Traffic Steering (RRTS). The RR and its route reflection clients form a RRTS domain. The RR is a centralized controller that is responsible for the BGP route decision of the whole domain. All other routers in the domain are as route reflection clients of the RR, each router will establish an I-BGP session to the RR, and there is no direct BGP sessions among these routers.

This looks no different from the current Route Reflector (RR) based architecture. For each client, it will still run as current, when received BGP routes from outside, it will transparently distribute the routes to the RR. For each route, the RR will make the decision for each relevant router and then install/distribute the route to each related router.

For example, for a path from Source 1 (S1) to Destination 1 (D1), if the computed path is: S1-A-B-C-D1, then the RR will distribute a route (D1) to C with the nexthop set to D1; a route (D1) to B with the nexthop set to C, and a route (D1) to A with the nexthop set to B, and finally the route (D1) will be distributed to S1 by A.

RRTS will not require the clients to make any changes. All the changes are made on the RR, the RR can apply any route or traffic engineering algorithms.

4. IANA Considerations

This document makes no request of IANA.

5. Acknowledgements

The authors would like to thank Bai Tao, Fengqing Yu for their contribution to this document.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

6.2. Informative References

[I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder,
"Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-08 (work in progress), December 2012.

[I-D.ietf-idr-bgp-optimal-route-reflection]
Raszuk, R., Cassar, C., Aman, E., Decraene, B., and S.
Litkowski, "BGP Optimal Route Reflection (BGP-ORR)",
draft-ietf-idr-bgp-optimal-route-reflection-05 (work in
progress), June 2013.

Authors' Addresses

Mach(Guoyi) Chen
Huawei Technologies

Email: mach.chen@huawei.com

Shunwan Zhuang
Huawei Technologies

Email: zhuangshunwan@huawei.com

Yongqing Zhu
China Telecom Co.,Ltd
109 West Zhongshan Ave,Tianhe District
Guangzhou 510630
China

Email: zhuyq@gsta.com

Subin Wang
China Telecom Co.,Ltd
109 West Zhongshan Ave,Tianhe District
Guangzhou 510630
China

Email: wangsb@gsta.com

GROW Working Group
Internet-Draft
Intended status: Informational
Expires: December 09, 2013

N. Hilliard
INEX
E. Jasinska
Microsoft Corporation
R. Raszuk
NTT I3
N. Bakker
AMS-IX B.V.
June 07, 2013

Internet Exchange Route Server Operations
draft-hilliard-ix-bgp-route-server-operations-03

Abstract

The popularity of Internet exchange points (IXPs) brings new challenges to interconnecting networks. While bilateral eBGP sessions between exchange participants were historically the most common means of exchanging reachability information over an IXP, the overhead associated with this interconnection method causes serious operational and administrative scaling problems for IXP participants.

Multilateral interconnection using Internet route servers can dramatically reduce the administrative and operational overhead of IXP participation and these systems used by many IXP participants as a preferred means of exchanging routing information.

This document describes operational considerations for multilateral interconnections at IXPs.

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 December 09, 2013.

Copyright Notice

Copyright (c) 2013 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 |
| 1.1. Notational Conventions | 3 |
| 2. Bilateral BGP Sessions | 3 |
| 3. Multilateral Interconnection | 4 |
| 4. Operational Considerations for Route Server Installations . . | 5 |
| 4.1. Path Hiding | 5 |
| 4.2. Route Server Scaling | 6 |
| 4.2.1. Tackling Scaling Issues | 6 |
| 4.2.1.1. View Merging and Decomposition | 6 |
| 4.2.1.2. Destination Splitting | 7 |
| 4.2.1.3. NEXT_HOP Resolution | 8 |
| 4.3. Prefix Leakage Mitigation | 8 |
| 4.4. Route Server Redundancy | 8 |
| 4.5. AS_PATH Consistency Check | 9 |
| 4.6. Export Routing Policies | 9 |
| 4.6.1. BGP Communities | 9 |
| 4.6.2. Internet Routing Registry | 9 |
| 4.6.3. Client-accessible Databases | 10 |
| 4.7. Layer 2 Reachability Problems | 10 |
| 5. Security Considerations | 10 |
| 6. IANA Considerations | 10 |
| 7. Acknowledgments | 11 |
| 8. References | 11 |
| 8.1. Normative References | 11 |
| 8.2. Informative References | 11 |
| Authors' Addresses | 12 |

1. Introduction

Internet exchange points (IXPs) provide IP data interconnection facilities for their participants, typically using shared Layer-2

networking media such as Ethernet. The Border Gateway Protocol (BGP) [RFC4271] is normally used to facilitate exchange of network reachability information over these media.

As bilateral interconnection between IXP participants requires operational and administrative overhead, BGP route servers [I-D.ietf-idr-ix-bgp-route-server] are often deployed by IXP operators to provide a simple and convenient means of interconnecting IXP participants with each other. A route server redistributes prefixes received from its BGP clients to other clients according to a pre-specified policy, and it can be viewed as similar to an eBGP equivalent of an iBGP [RFC4456] route reflector.

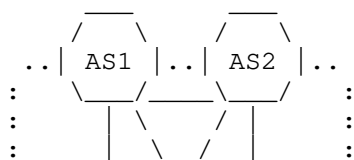
Route servers at IXPs require careful management and it is important for route server operators to thoroughly understand both how they work and what their limitations are. In this document, we discuss several issues of operational relevance to route server operators and provide recommendations to help route server operators provision a reliable interconnection service.

1.1. Notational Conventions

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 [RFC2119].

2. Bilateral BGP Sessions

Bilateral interconnection is a method of interconnecting routers using individual BGP sessions between each participant router on an IXP, in order to exchange reachability information. If an IXP participant wishes to implement an open interconnection policy - i.e. a policy of interconnecting with as many other IXP participants as possible - it is necessary for the participant to liaise with each of their intended interconnection partners. Interconnection can then be implemented bilaterally by configuring a BGP session on both participants' routers to exchange network reachability information. If each exchange participant interconnects with each other participant, a full mesh of BGP sessions is needed, as shown in Figure 1.



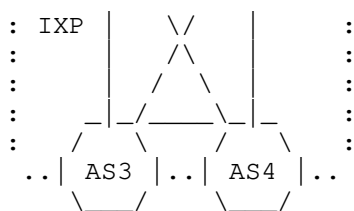


Figure 1: Full-Mesh Interconnection at an IXP

Figure 1 depicts an IXP platform with four connected routers, administered by four separate exchange participants, each of them with a locally unique autonomous system number: AS1, AS2, AS3 and AS4. Each of these four participants wishes to exchange traffic with all other participants; this is accomplished by configuring a full mesh of BGP sessions on each router connected to the exchange, resulting in 6 BGP sessions across the IXP fabric.

The number of BGP sessions at an exchange has an upper bound of $n*(n-1)/2$, where n is the number of routers at the exchange. As many exchanges have large numbers of participating networks, the amount of administrative and operation overhead required to implement an open interconnection scales quadratically. New participants to an IXP require significant initial resourcing in order to gain value from their IXP connection, while existing exchange participants need to commit ongoing resources in order to benefit from interconnecting with these new participants.

3. Multilateral Interconnection

Multilateral interconnection is implemented using a route server configured to use BGP to distribute network layer reachability information (NLRI) among all client routers. The route server preserves the BGP NEXT_HOP attribute from all received NLRI UPDATE messages, and passes these messages with unchanged NEXT_HOP to its route server clients, according to its configured routing policy, as described in [I-D.ietf-idr-ix-bgp-route-server]. Using this method of exchanging NLRI messages, an IXP participant router can receive an aggregated list of prefixes from all other route server clients using a single BGP session to the route server instead of depending on BGP sessions with each other router at the exchange. This reduces the overall number of BGP sessions at an Internet exchange from $n*(n-1)/2$ to n , where n is the number of routers at the exchange.

Although a route server uses BGP to exchange reachability information with each of its clients, it does not forward traffic itself and is therefore not a router.

In practical terms, this allows dense interconnection between IXP participants with low administrative overhead and significantly simpler and smaller router configurations. In particular, new IXP participants benefit from immediate and extensive interconnection, while existing route server participants receive reachability information from these new participants without necessarily having to modify their configurations.

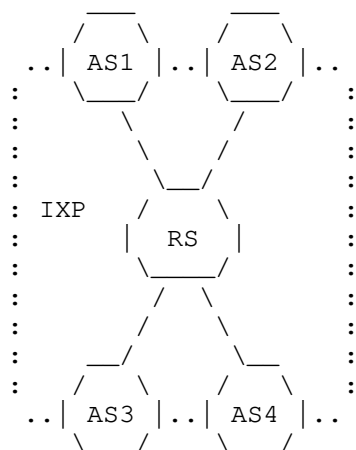


Figure 2: IXP-based Interconnection with Route Server

As illustrated in Figure 2, each router on the IXP fabric requires only a single BGP session to the route server, from which it can receive reachability information for all other routers on the IXP which also connect to the route server.

4. Operational Considerations for Route Server Installations

4.1. Path Hiding

"Path hiding" is a term used in [I-D.ietf-idr-ix-bgp-route-server] to describe the process whereby a route server may mask individual paths by applying conflicting routing policies to its Loc-RIB. When this happens, route server clients receive incomplete information from the route server about network reachability.

There are several approaches which may be used to mitigate against the effect of path hiding; these are described in [I-D.ietf-idr-ix-bgp-route-server]. However, the only method which does not require explicit support from the route server client is for the route server itself to maintain a individual Loc-RIB for each client which is the subject of conflicting routing policies.

4.2. Route Server Scaling

While deployment of multiple Loc-RIBs on the route server presents a simple way to avoid the path hiding problem noted in Section 4.1, this approach requires significantly more computing resources on the route server than where a single Loc-RIB is deployed for all clients. As the [RFC4271] BGP decision process must be applied to all Loc-RIBs deployed on the route server, both CPU and memory requirements on the host computer scale approximately according to $O(P * N)$, where P is the total number of unique paths received by the route server and N is the number of route server clients which require a unique Loc-RIB. As this is a super-linear scaling relationship, large route servers may derive benefit from deploying per-client Loc-RIBs only where they are required.

Regardless of any Loc-RIB optimization technique is implemented, the route server's control plane bandwidth requirements will scale according to $O(P * N)$, where P is the total number of unique paths received by the route server and N is the total number of route server clients. In the case where P_{avg} (the arithmetic mean number of unique paths received per route server client) remains roughly constant even as the number of connected clients increases, this relationship can be rewritten as $O((P_{avg} * N) * N)$ or $O(N^2)$. This quadratic upper bound on the network traffic requirements indicates that the route server model will not scale to arbitrarily large sizes.

This scaling analysis presents problems in three key areas: route processor CPU overhead associated with BGP decision process calculations, the memory requirements for handling many different BGP path entries, and the network traffic bandwidth required to distribute these prefixes from the route server to each route server client.

4.2.1. Tackling Scaling Issues

The network traffic scaling issue presents significant difficulties with no clear solution - ultimately, each client must receive a UPDATE for each unique prefix received by the route server. However, there are several potential methods for dealing with the CPU and memory resource requirements of route servers.

4.2.1.1. View Merging and Decomposition

View merging and decomposition, outlined in [RS-ARCH], describes a method of optimising memory and CPU requirements where multiple route server clients are subject to exactly the same routing policies. In this situation, the multiple Loc-RIB views required by each client are merged into a single view.

There are several variations of this approach. If the route server operator has prior knowledge of interconnection relationships between route server clients, then the operator may configure separate Loc-RIBs only for route server clients with unique outbound routing policies. As this approach requires prior knowledge of interconnection relationships, the route server operator must depend on each client sharing their interconnection policies, either in a internal provisioning database controlled by the operator, or else in an external data store such as an Internet Routing Registry Database.

Conversely, the route server implementation itself may implement internal view decomposition by creating virtual Loc-RIBs based on a single in-memory master Loc-RIB, with delta differences for each prefix subject to different routing policies. This allows a more granular and flexible approach to the problem of Loc-RIB scaling, at the expense of requiring a more complex in-memory Loc-RIB structure.

Whatever method of view merging and decomposition is chosen on a route server, pathological edge cases can be created whereby they will scale no better than fully non-optimised per-client Loc-RIBs. However, as most route server clients connect to a route server for the purposes of reducing overhead, rather than implementing complex per-client routing policies, edge cases tend not to arise in practice.

4.2.1.2. Destination Splitting

Destination splitting, also described in [RS-ARCH], describes a method for route server clients to connect to multiple route servers and to send non-overlapping sets of prefixes to each route server. As each route server computes the best path for its own set of prefixes, the quadratic scaling requirement operates on multiple smaller sets of prefixes. This reduces the overall computational and memory requirements for managing multiple Loc-RIBs and performing the best-path calculation on each. In order for this method to perform well, destination splitting would require significant co-ordination between the route server operator and each route server client. In practice, this level of close co-ordination between IXP operators and their participants tends not to occur, suggesting that the approach is unlikely to be of any real use on production IXPs.

4.2.1.3. NEXT_HOP Resolution

As route servers are usually deployed at IXPs which use flat layer 2 networks, recursive resolution of the NEXT_HOP attribute is generally not required, and can be replaced by a simple check to ensure that the NEXT_HOP value for each prefix is a network address on the IXP LAN's IP address range.

4.3. Prefix Leakage Mitigation

Prefix leakage occurs when a BGP client unintentionally distributes NLRI UPDATE messages to one or more neighboring BGP routers. Prefix leakage of this form to a route server can cause serious connectivity problems at an IXP if each route server client is configured to accept all prefix UPDATE messages from the route server. It is therefore RECOMMENDED when deploying route servers that, due to the potential for collateral damage caused by NLRI leakage, route server operators deploy prefix leakage mitigation measures in order to prevent unintentional prefix announcements or else limit the scale of any such leak. Although not foolproof, per-client inbound prefix limits can restrict the damage caused by prefix leakage in many cases. Per-client inbound prefix filtering on the route server is a more deterministic and usually more reliable means of preventing prefix leakage, but requires more administrative resources to maintain properly.

If a route server operator implements per-client inbound prefix filtering, then it is RECOMMENDED that the operator also builds in mechanisms to automatically compare the Adj-RIB-In received from each client with the inbound prefix lists configured for those clients. Naturally, it is the responsibility of the route server client to ensure that their stated prefix list is compatible with what they announce to an IXP route server. However, many network operators do not carefully manage their published routing policies and it is not uncommon to see significant variation between the two sets of prefixes. Route server operator visibility into this discrepancy can provide significant advantages to both operator and client.

4.4. Route Server Redundancy

As the purpose of an IXP route server implementation is to provide a reliable reachability brokerage service, it is RECOMMENDED that exchange operators who implement route server systems provision multiple route servers on each shared Layer-2 domain. There is no requirement to use the same BGP implementation or operating system for each route server on the IXP fabric; however, it is RECOMMENDED that where an operator provisions more than a single server on the same shared Layer-2 domain, each route server implementation be

configured equivalently and in such a manner that the path reachability information from each system is identical.

4.5. AS_PATH Consistency Check

[RFC4271] requires that every BGP speaker which advertises a route to another external BGP speaker prepends its own AS number as the last element of the AS_PATH sequence. Therefore the leftmost AS in an AS_PATH attribute should be equal to the autonomous system number of the BGP speaker which sent the UPDATE message.

As [I-D.ietf-idr-ix-bgp-route-server] suggests that route servers should not modify the AS_PATH attribute, a consistency check on the AS_PATH of an UPDATE received by a route server client would normally fail. It is therefore RECOMMENDED that route server clients disable the AS_PATH consistency check towards the route server.

4.6. Export Routing Policies

Policy filtering is commonly implemented on route servers to provide prefix distribution control mechanisms for route server clients. A route server "export" policy is a policy which affects prefixes sent from the route server to a route server client. Several different strategies are commonly used for implementing route server export policies.

4.6.1. BGP Communities

Prefixes sent to the route server are tagged with specific [RFC1997] or [RFC4360] BGP community attributes, based on pre-defined values agreed between the operator and all client. Based on these community tags, prefixes may be propagated to all other clients, a subset of clients, or none. This mechanism allows route server clients to instruct the route server to implement per-client export routing policies.

As both standard and extended BGP communities values are restricted to 6 octets, the route server operator should take care to ensure that the predefined BGP community values mechanism used on their route server is compatible with [RFC4893] 4-octet autonomous system numbers.

4.6.2. Internet Routing Registry

Internet Routing Registry databases (IRRDBs) may be used by route server operators to implement construct per-client routing policies. [RFC2622] Routing Policy Specification Language (RPSL) provides an comprehensive grammar for describing interconnection relationships,

and several toolsets exist which can be used to translate RPSL policy description into route server configurations.

4.6.3. Client-accessible Databases

Should the route server operator not wish to use either BGP community tags or the public IRRDBs for implementing client export policies, they may implement their own routing policy database system for managing their clients' requirements. A database of this form SHOULD allow a route server client operator to update their routing policy and provide a mechanism for allowing the client to specify whether they wish to exchange all their prefixes with any other route server client. Optionally, the implementation may allow a client to specify unique routing policies for individual prefixes over which they have routing policy control.

4.7. Layer 2 Reachability Problems

Layer 2 reachability problems on an IXP can cause serious operational problems for IXP participants which depend on route servers for interconnection. Ethernet switch forwarding bugs have occasionally been observed to cause non-commutative reachability. For example, given a route server and two IXP participants, A and B, if the two participants can reach the route server but cannot reach each other, then traffic between the participants may be dropped until such time as the layer 2 forwarding problem is resolved. This situation does not tend to occur in bilateral interconnection arrangements, as the routing control path between the two hosts is usually (but not always, due to IXP inter-switch connectivity load balancing algorithms) the same as the data path between them.

Problems of this form can be dealt with using [RFC5881] bidirectional forwarding detection. However, as this is a bilateral protocol configured between routers, and as there is currently no means for automatic configuration of BFD between route server clients, BFD does not currently provide an optimal means of handling the problem.

5. Security Considerations

On route server installations which do not employ path hiding mitigation techniques, the path hiding problem outlined in section Section 4.1 can be used in certain circumstances to proactively block third party prefix announcements from other route server clients.

6. IANA Considerations

There are no IANA considerations.

7. Acknowledgments

The authors would like to thank Chris Hall, Ryan Bickhart and Steven Bakker for their valuable input.

In addition, the authors would like to acknowledge the developers of BIRD, OpenBGPD and Quagga, whose open source BGP implementations include route server capabilities which are compliant with this document.

8. References

8.1. Normative References

- [I-D.ietf-idr-ix-bgp-route-server]
Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker,
"Internet Exchange Route Server", draft-ietf-idr-ix-bgp-
route-server-02 (work in progress), February 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.

8.2. Informative References

- [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP
Communities Attribute", RFC 1997, August 1996.
- [RFC2622] Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D.,
Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra,
"Routing Policy Specification Language (RPSL)", RFC 2622,
June 1999.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway
Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
Communities Attribute", RFC 4360, February 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route
Reflection: An Alternative to Full Mesh Internal BGP
(IBGP)", RFC 4456, April 2006.
- [RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS
Number Space", RFC 4893, May 2007.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an
IANA Considerations Section in RFCs", BCP 26, RFC 5226,
May 2008.

- [RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.
- [RS-ARCH] Govindan, R., Alaettinoglu, C., Varadhan, K., and D. Estrin, "A Route Server Architecture for Inter-Domain Routing", 1995,
<<http://www.cs.usc.edu/research/95-603.ps.Z>>.

Authors' Addresses

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

Elisa Jasinska
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
US

Email: ejas@microsoft.com

Robert Raszuk
NTT I3
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

Niels Bakker
AMS-IX B.V.
Westeinde 12
Amsterdam, NH 1017 ZN
NL

Email: niels.bakker@ams-ix.net

Internet Engineering Task Force
Internet-Draft
Intended status: Best Current Practice
Expires: June 2, 2015

J. Durand
CISCO Systems, Inc.
I. Pepelnjak
NIL
G. Doering
SpaceNet
December 2, 2014

BGP operations and security
draft-ietf-opsec-bgp-security-07.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 is 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, TCP-AO, 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.

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

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 May 29, 2015.

Copyright Notice

Copyright (c) 2014 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 | 3 |
| 2. Scope of the document | 3 |
| 3. Definitions and Acronyms | 4 |
| 4. Protection of the BGP speaker | 4 |
| 5. Protection of BGP sessions | 5 |
| 5.1. Protection of TCP sessions used by BGP | 5 |
| 5.2. BGP TTL security (GTSM) | 6 |
| 6. Prefix filtering | 6 |
| 6.1. Definition of prefix filters | 6 |
| 6.1.1. Special purpose prefixes | 6 |
| 6.1.2. Prefixes not allocated | 7 |
| 6.1.3. Prefixes too specific | 11 |
| 6.1.4. Filtering prefixes belonging to the local AS and downstreams | 11 |
| 6.1.5. IXP LAN prefixes | 11 |
| 6.1.6. The default route | 12 |
| 6.2. Prefix filtering recommendations in full routing networks | 13 |
| 6.2.1. Filters with Internet peers | 13 |
| 6.2.2. Filters with customers | 15 |
| 6.2.3. Filters with upstream providers | 15 |
| 6.3. Prefix filtering recommendations for leaf networks . . . | 16 |
| 6.3.1. Inbound filtering | 16 |
| 6.3.2. Outbound filtering | 16 |
| 7. BGP route flap dampening | 17 |
| 8. Maximum prefixes on a peering | 17 |
| 9. AS-path filtering | 17 |
| 10. Next-Hop Filtering | 19 |
| 11. BGP community scrubbing | 20 |
| 12. Change logs | 20 |
| 12.1. Diffs between draft-jdurand-bgp-security-01 and draft- jdurand-bgp-security-00 | 20 |

| | | |
|--------------------|---|----|
| 12.2. | Diffs between draft-jdurand-bgp-security-02 and draft-jdurand-bgp-security-01 | 21 |
| 12.3. | Diffs between draft-ietf-opsec-bgp-security-00 and draft-jdurand-bgp-security-02 | 22 |
| 12.4. | Diffs between draft-ietf-opsec-bgp-security-01 and draft-ietf-opsec-bgp-security-00 | 22 |
| 12.5. | Diffs between draft-ietf-opsec-bgp-security-02 and draft-ietf-opsec-bgp-security-01 | 23 |
| 12.6. | Diffs between draft-ietf-opsec-bgp-security-03 and draft-ietf-opsec-bgp-security-02 | 24 |
| 12.7. | Diffs between draft-ietf-opsec-bgp-security-04 and draft-ietf-opsec-bgp-security-03 | 25 |
| 12.8. | Diffs between draft-ietf-opsec-bgp-security-05 and draft-ietf-opsec-bgp-security-04 | 25 |
| 12.9. | Diffs between draft-ietf-opsec-bgp-security-06 and draft-ietf-opsec-bgp-security-05 | 25 |
| 13. | Acknowledgements | 26 |
| 14. | IANA Considerations | 26 |
| 15. | Security Considerations | 26 |
| 16. | References | 27 |
| 16.1. | Normative References | 27 |
| 16.2. | Informative References | 27 |
| Appendix A. | IXP LAN prefix filtering - example | 29 |
| Authors' Addresses | | 30 |

1. Introduction

BGP (Border Gateway Protocol - RFC 4271 [2]) is the protocol used in the Internet to exchange routing information between network domains. BGP does not directly include mechanisms that control that routes exchanged conform to the various guidelines defined by the Internet community. This document intends to both summarize common existing guidelines and help network administrators apply coherent BGP policies.

2. Scope of the document

The guidelines defined in this document are intended for generic Internet BGP peerings. The nature of the Internet is such that Autonomous Systems can always agree on exceptions to a common framework for relevant local needs, and therefore configure a BGP session in a manner that may differ from the recommendations provided in this document. While this is perfectly acceptable, every configured exception might have an impact on the entire inter-domain routing environment and network administrators SHOULD carefully appraise this impact before implementation.

3. Definitions and Accronyms

- o ACL: Access Control List
- o ASN: Autonomous System Number
- o IRR: Internet Routing Registry
- o IXP: Internet eXchange Point
- o LIR: Local Internet Registry
- o pMTUd: Path MTU Discovery
- o RIR: Regional Internet Registry
- o Tier 1 transit provider: an IP transit provider which can reach any network on the Internet without purchasing transit services
- o uRPF: Unicast Reverse Path Forwarding

4. Protection of the BGP speaker

The BGP speaker needs to be protected from attempts to subvert the BGP session. This protection SHOULD be achieved by an Access Control List (ACL) which would discard all packets directed to TCP port 179 on the local device and sourced from an address not known or permitted to become a BGP neighbor. Experience has shown that natural protection TCP should offer is not always sufficient as it is sometimes run in control-plane software: in the absence of ACLs it is possible to attack a BGP speaker by simply sending a high volume of connection requests to it.

If supported, an ACL specific to the control-plane of the router SHOULD be used (receive-ACL, control-plane policing, etc.), to avoid configuration of data-plane filters for packets transiting through the router (and therefore not reaching the control plane). If the hardware can not do that, interface ACLs can be used to block packets addressed 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.

In addition to strict filtering, rate-limiting MAY be configured for accepted BGP traffic. Rate-limiting BGP traffic consists in permitting only a certain quantity of bits per second (or packets per second) of BGP traffic to the control plane. This protects the BGP

router control plane in case the amount of BGP traffic overcomes platform capabilities.

Filtering and rate-limiting of control-plane traffic is a wider topic than "just for BGP" (if network administrator brings down a router by overloading one of the other protocols from remote, BGP is harmed as well). For a more detailed recommendation on how to protect the router's control plane, see RFC 6192 [11].

5. Protection of BGP sessions

Current security issues of TCP-based protocols (therefore including BGP) have been documented in RFC 6952 [14]. The following subsections list the major points raised in this RFC and give best practices related to TCP session protection for BGP operation.

5.1. Protection of TCP sessions used by BGP

Attacks on TCP sessions used by BGP (aka BGP sessions), for example sending spoofed TCP RST packets, could bring down a BGP peering. 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).

BGP sessions can be secured with a variety of mechanisms. MD5 protection of TCP session header, described in RFC 2385 [7], was the first such mechanism. It is now deprecated by TCP Authentication Option (TCP-AO, RFC 5925 [4]) which offers stronger protection. While MD5 is still the most used mechanism due to its availability in vendor's equipment, TCP-AO SHOULD be preferred when implemented.

IPsec could also be used for session protection. At the time this document is published, there is not enough experience on impacts of the use of IPsec for BGP peerings and further analysis is required to define guidelines.

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 NOT REQUIRED even when peerings are established over shared networks where spoofing can be done (like IXPs), but operators are RECOMMENDED to consider the trade-offs and to apply TCP session protection where appropriate.

Network administrators SHOULD block spoofed packets (packets with a source IP address belonging to their IP address space) at all edges of their network (see RFC 2827 [8] and RFC 3704 [9]). This protects

the TCP session used by iBGP from attackers outside the Autonomous System.

5.2. BGP TTL security (GTSM)

BGP sessions can be made harder to spoof with the Generalized TTL Security Mechanisms (GTSM, aka TTL security), defined in RFC 5082 [3]. Instead of sending TCP packets with TTL value of 1, the BGP speakers send the TCP packets with TTL value of 255 and the receiver checks that the TTL value equals 255. Since it's impossible to send an IP packet with TTL of 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. Network administrators SHOULD implement TTL security on directly connected BGP peerings.

GTSM could also be applied to multi-hop BGP peering as well. To achieve this TTL needs to be configured with proper value depending on the distance between BGP speakers (using principle described above). Nevertheless it is not as effective as anyone inside the TTL diameter could spoof the TTL.

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

6. 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 6), AS paths (as-path filters, Section 9) or any other attributes of a BGP prefix (for example, BGP communities, Section 11).

6.1. Definition of prefix filters

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

6.1.1. Special purpose prefixes

6.1.1.1. IPv4 special purpose prefixes

IANA IPv4 Special-Purpose Address Registry [22] maintains the list of IPv4 special purpose prefixes and their routing scope, and SHOULD be used for prefix filters configuration. Prefixes with value "False" in column "Global" SHOULD be discarded on Internet BGP peerings.

6.1.1.2. IPv6 special purpose prefixes

IANA IPv6 Special-Purpose Address Registry [23] maintains the list of IPv6 special purpose prefixes and their routing scope, and SHOULD be used for prefix filters configuration. Only prefixes with value "False" in column "Global" SHOULD be discarded on Internet BGP peerings.

6.1.2. Prefixes not allocated

IANA allocates prefixes to RIRs which in turn allocate prefixes to LIRs (Local Internet Registries). It is wise not to accept routing table prefixes that are not allocated by IANA and/or RIRs. This section details the options for building a list of allocated prefixes at every level. It is important to understand that filtering prefixes not allocated requires constant updates as prefixes are continually allocated. Therefore automation of such prefix filters is key for the success of this approach. Network administrators SHOULD NOT consider solutions described in this section if they are not capable of maintaining updated prefix filters: the damage would probably be worse than the intended security policy.

6.1.2.1. IANA allocated prefix filters

IANA has allocated all the IPv4 available space. Therefore there is no reason why network administrators would keep checking that prefixes they receive from BGP peers are in the IANA allocated IPv4 address space [24]. No specific filters need to be put in place by administrators who want to make sure that IPv4 prefixes they receive in BGP updates 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 [25]. As IANA keeps allocating prefixes to RIRs, the aforementioned list should be checked regularly against changes and if they occur, prefix filters should be computed and pushed on network devices. The list could also be pulled directly by routers when they implement such mechanisms. 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 for doing this step quickly and frequently. However, network administrators SHOULD ensure that all IPv6 prefix filters are updated within maximum one month after any change in the list of IPv6 prefix allocated by IANA.

If process in place (manual or automatic) cannot guarantee that the list is updated regularly then it's better not to configure any

filters 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 an extra work for RIRs that had to "de-bogonize" the newly allocated prefixes.

6.1.2.2. RIR allocated prefix filters

A more precise check can be performed when one would like to make sure that prefixes they receive are being originated or transited by autonomous systems entitled to do so. It has been observed in the past that an AS (Autonomous System) could easily advertise someone else's prefix (or more specific prefixes) and create black holes or security threats. To partially mitigate this 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.

6.1.2.2.1. 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 - RFC 4012 [10]. Network administrators are given privileges to describe routing policies of their own networks in the IRR and information is published, usually publicly. A majority of Regional Internet Registries do also operate an IRR and can control that registered routes conform to prefixes allocated or directly assigned. However, it should be noted that the list of such prefixes is not necessarily a complete list, and as such the list of routes in an IRR is not the same as the set of RIR allocated prefixes.

It is possible to use the IRR information to build, for a given neighbor autonomous system, a list of prefixes originated or transited which one may accept. 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-SETs. 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 likely to announce. For each of these AS numbers, it is also easy to check in the corresponding IRR for all associated prefixes. With these two mechanisms a script can build for a given peer the list of allowed prefixes and the AS number from which they should be originated. One could decide not use the origin information and only build monolithic prefix filters from fetched data.

As prefixes, AS numbers and AS-SETs may not all be under the same RIR authority, a difficulty resides choosing for each object the appropriate IRR to poll. Some IRRs 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 a popular IRR containing many sources (such as RADB [26], the Routing Assets Database) and only select as sources some desired RIRs and trusted major ISPs (Internet Service Providers).

As objects in IRRs may frequently vary over time, it is important that prefix filters computed using this mechanism are refreshed regularly. A daily basis could even be 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 tens of thousands configuration lines for some important peers. To manage this complexity, network administrators could for example use IRRToolSet [29], a set of tools making it possible to simplify the creation of automated filters configuration from policies stored in IRR.

Last but not least, network administrators SHOULD publish and maintain their resources properly in IRR database maintained by their RIR, when available.

6.1.2.2.2. SIDR - Secure Inter Domain Routing

An infrastructure called SIDR (Secure Inter-Domain Routing), described in RFC 6480 [12] has been designed to secure Internet advertisements. At the time this document is written, many documents have been published and a framework with a complete set of protocols is proposed so that advertisements can be checked against signed routing objects in RIR routing registries. There are basically two services that SIDR offers:

- o Origin validation, described in RFC 6811 [5], seeks at making sure that attributes associated with a routes are correct (the major point being the validation of the AS number originating this route). Origin validation is now operational (Internet registries, protocols, implementations on some routers...) and in theory it can be implemented knowing that the proportion of signed resources is still low at the time this document is written.
- o Path validation provided by BGPsec [27] seeks at making sure that no ones announce fake/wrong BGP paths that would attract traffic for a given destination, see RFC 7132 [16]. BGPsec is still an on-going work item at the time this document is written and therefore cannot be implemented.

Implementing SIDR mechanisms is expected to solve many of BGP routing security problems in the long term but it may take time for deployments to be made and objects to become signed. It also has to be pointed that SIDR infrastructure is complementing (not replacing) the security best practices listed in this document. Network administrators SHOULD therefore implement any SIDR proposed mechanism (example: route origin validation) on top of the other existing mechanisms even if they could sometimes appear targeting the same goal.

If route origin validation is implemented, reader SHOULD refer to rules described in RFC 7115 [15]. In short, each external route received on a router SHOULD be checked against the RPKI data set:

- o If a corresponding ROA (Route Origin Authorization) is found and is valid then the prefix SHOULD be accepted.
- o If the ROA is found and is INVALID then the prefix SHOULD be discarded.
- o If an ROA is not found then the prefix SHOULD be accepted but corresponding route SHOULD be given a low preference.

In addition to this, network administrators SHOULD sign their routing objects so their routes can be validated by other networks running origin validation.

One should understand that the RPKI model brings new interesting challenges. The paper On the Risk of Misbehaving RPKI Authorities [30] explains how RPKI model can impact the Internet if authorities don't behave as they are supposed to do. Further analysis is certainly required on RPKI, which carries part of BGP security.

6.1.3. 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 the RIPE community has documented that as of the time of writing of this document, IPv4 prefixes longer than /24 and IPv6 prefixes longer than /48 are generally not announced/accepted in the Internet [19] [20]. These values may change in the future.

6.1.4. Filtering prefixes belonging to the local AS and downstreams

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) from leaking 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.

In some cases, for example in multi-homing scenarios, such filters SHOULD NOT be applied as this would break the desired redundancy.

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.

6.1.5. IXP LAN prefixes

6.1.5.1. Network security

When a network is present on an IXP and peers with other IXP members over a common subnet (IXP LAN prefix), it SHOULD NOT accept more specific prefixes for the IXP LAN prefix from any of its external BGP peers. Accepting these routes may 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 SHOULD 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.

6.1.5.2. pMTUd and the 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 6.1.2.2.1 as well as "prefixes too specific" filters described in Section 6.1.3. 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.

Appendix Appendix A gives an example of guidelines regarding IXP LAN prefix.

6.1.6. The default route

6.1.6.1. IPv4

The 0.0.0.0/0 prefix is likely not intended to be accepted nor advertised other than in specific customer / provider configurations, general filtering outside of these is RECOMMENDED.

6.1.6.2. IPv6

The `::/0` prefix is likely not intended to be accepted nor advertised other than in specific customer / provider configurations, general filtering outside of these is RECOMMENDED.

6.2. 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 and uses prefix filters defined in previous section Section 6.1.

6.2.1. Filters with Internet peers

6.2.1.1. 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.

6.2.1.1.1. Inbound filtering loose option

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

- o Prefixes not globally routable (Section 6.1.1)
- o Prefixes not allocated by IANA (IPv6 only) (Section 6.1.2.1)
- o Routes too specific (Section 6.1.3)
- o Prefixes belonging to the local AS (Section 6.1.4)
- o IXP LAN prefixes (Section 6.1.5)
- o The default route (Section 6.1.6)

6.2.1.1.2. Inbound filtering strict option

In this case, filters are applied to make sure advertisements strictly conform to what is declared in routing registries (Section 6.1.2.2). Warning is given as registries are not always

accurate (prefixes missing, wrong information...) This varies across the registries and regions of the Internet. Before applying a strict policy the reader SHOULD check the impact on the filter and make sure solution is not worse than the problem.

Also in case of script failure each administrator may decide if all routes are accepted or rejected depending on routing policy. While accepting the routes during that time frame could break the BGP routing security, rejecting them might re-route too much traffic on transit peers, and could cause more harm than what a loose policy would have done.

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

- o Prefixes not globally routable (Section 6.1.1)
- o Routes too specific (Section 6.1.3)
- o Prefixes belonging to the local AS (Section 6.1.4)
- o IXP LAN prefixes (Section 6.1.5)
- o The default route (Section 6.1.6)

6.2.1.2. Outbound filtering

Configuration should be put in place to make sure that only appropriate prefixes are sent. These can be, for example, prefixes belonging to both the network in question and its downstreams. This can be achieved by using a combination of BGP communities, AS-paths or both. It can also be desirable that following filters are positioned before to avoid unwanted route announcement due to bad configuration:

- o Prefixes not globally routable (Section 6.1.1)
- o Routes too specific (Section 6.1.3)
- o IXP LAN prefixes (Section 6.1.5)
- o The default route (Section 6.1.6)

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.

6.2.2. Filters with customers

6.2.2.1. Inbound filtering

The inbound policy with end customers is pretty straightforward: only customers prefixes SHOULD 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.

The 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 6.2.1.1 can be applied.

6.2.2.2. Outbound filtering

The outbound policy with customers may vary according to the routes customer wants to receive. In the simplest possible scenario, the customer may only want 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 of the Internet table), the following filters can be simply applied on the BGP peering:

- o Prefixes not globally routable (Section 6.1.1)
- o Routes too specific (Section 6.1.3)
- o The default route (Section 6.1.6)

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, network administrators may decide to originate it only for peerings where it has to be advertised.

6.2.3. Filters with upstream providers

6.2.3.1. Inbound filtering

In case the full routing table is desired from the upstream, the prefix filtering to apply is the same as the one for peers Section 6.2.1.1 with the exception of the default route. The default route can be desired from an upstream provider in addition to 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.

6.2.3.2. Outbound filtering

The filters to be applied would most likely not differ much from the ones applied for Internet peers (Section 6.2.1.2). But different policies could be applied in case it is desired that a particular upstream does not provide transit to all the prefixes.

6.3. Prefix filtering recommendations for leaf networks

6.3.1. Inbound filtering

The leaf network will deploy the filters corresponding to the routes it is requesting from its upstream. In case a default route is requested, a simple inbound filter can be applied to accept only the default route (Section 6.1.6). In case the leaf network is not capable of listing the prefixes 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:

- o Prefixes not routable (Section 6.1.1)
- o Routes too specific (Section 6.1.3)
- o Prefixes belonging to local AS (Section 6.1.4)
- o The default route (Section 6.1.6) depending if the route is requested or not

6.3.2. 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 6.2.1.2 to avoid announcing invalid routes to its upstream provider.

7. BGP route flap dampening

The 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. Studies have shown that implementations of BGP route flap dampening could cause more harm than they solve problems and therefore RIPE community has in the past recommended not using BGP route flap dampening [18]. Studies have then been conducted to propose new route flap dampening thresholds in order to make the solution "usable", see RFC 7196 [6] and RIPE has reviewed its recommendations in [21]. This document RECOMMENDS following IETF and RIPE recommendations and only use BGP route flap dampening with the adjusted configured thresholds.

8. 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:

- o 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. Network administrators 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.
- o From upstreams which provide full routing, it is RECOMMENDED to have a limit 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 two 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.

9. AS-path filtering

This section lists the RECOMMENDED practices when processing BGP AS-paths:

- o Network administrators SHOULD accept from customers only AS(4)-Paths containing ASNs belonging to (or authorized to transit

through) the customer. If network administrators can not build and generate filtering expressions to implement this, they SHOULD consider accepting only path lengths relevant to the type of customer they have (as in, if these customers are a leaf or have customers of their own), and try to discourage excessive prepending in such paths. This loose policy could be combined with filters for specific AS(4)-Paths that must not be accepted if advertised by the customer, such as upstream transit provider or peer ASNs.

- o Network administrators SHOULD NOT accept prefixes with private AS numbers in the AS-path except from customers. Exception: an upstream offering some particular service like black-hole origination based on a private AS number. Customers should be informed by their upstream in order to put in place ad-hoc policy to use such services.
- o Network administrators SHOULD NOT accept prefixes when the first AS number in the AS-path is not the one of the peer unless the peering is done toward a BGP route-server [17] (for example on an IXP) with transparent AS path handling. In that case 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.
- o Network administrators SHOULD NOT advertise prefixes with non-empty AS-path unless they intend to be transit for these prefixes.
- o Network administrators SHOULD NOT advertise prefixes with upstream AS numbers in the AS-path to their peering AS unless they intend to be transit for these prefixes.
- o Private AS numbers are conventionally used in contexts that are "private" and SHOULD NOT be used in advertisements to BGP peers that are not party to such private arrangements, and should be stripped when received from BGP peers that are not party to such private arrangements.
- o Network administrators SHOULD NOT override BGP's default behavior accepting their own AS number in the AS-path. In case an exception to this is required, impacts should be studied carefully as this can create severe impact on routing.

AS-path filtering should be further analyzed when ASN renumbering is done. Such operation is common and mechanisms exist to allow smooth ASN migration [28]. The usual migration technique, local to a router, consists in modifying the AS-path so it is presented to a peer with the previous ASN, as if no renumbering was done. This

makes it possible to change ASN of a router without reconfiguring all eBGP peers at the same time (as this operation would require synchronization with all peers attached to that router). During this renumbering operation, rules described above may be adjusted.

10. Next-Hop Filtering

If peering on a shared network, like an IXP, 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 [17], 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, an inbound route policy SHOULD be applied on IXP peerings in order to set the next-hop for accepted prefixes to the BGP peer IP address (belonging to the IXP LAN) that sent the prefix (which is what "next-hop-self" would enforce on the sending side).

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

This policy also SHOULD be adjusted if Remote Triggered Black Holing best practice (aka RTBH - RFC 6666 [13]) is implemented. In that case network administrators would apply a well-known BGP next-hop for routes they want to filter (if an Internet threat is observed from/to this route for example). This well known next-hop will be statically routed to a null interface. In combination with unicast RPF check, this will discard traffic from and toward this prefix. Peers can exchange information about black-holes using for example particular BGP communities. Network administrators could propagate black-holes information to their peers using agreed BGP community: when receiving a route with that community a configured policy could change the next-hop in order to create the black hole.

11. BGP community scrubbing

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

- o Network administrators SHOULD scrub inbound communities with their number in the high-order bits, and allow only those communities that customers/peers can use as a signaling mechanism
- o Networks administrators SHOULD NOT remove other communities applied on received routes (communities not removed after application of previous statement). In particular they SHOULD keep original communities when they apply a community. Customers might need them to communicate with upstream providers. In particular network administrators SHOULD NOT (generally) remove the no-export community as it is usually announced by their peer for a certain purpose.

12. Change logs

!!! NOTE TO THE RFC EDITOR: THIS SECTION WAS ADDED TO TRACK CHANGES AND FACILITATE WORKING GROUP COLLABORATION. IT MUST BE DELETED BEFORE PUBLICATION !!!

12.1. 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:

- o "This documents" typo corrected in the former abstract
- o Add normative reference for RFC5082 in former section 3.2
- o "Non routable" changed in title of former section 4.1.1
- o Correction of typo for IPv4 loopback prefix in former section 4.1.1.1
- o Added shared transition space 100.64.0.0/10 in former section 4.1.1.1
- o Clarification that 2002::/16 6to4 prefix can cross network boundaries in former section 4.1.1.2
- o Rationale of 2000::/3 explained in former section 4.1.1.2

- o 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
 - o 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
 - o Replace RIR database with IRR. A definition of IRR is added in former section 4.1.2.2
 - o Remove any reference to anti-spoofing in former section 4.1.4
 - o Clarification for IXP LAN prefix and pMTUd problem in former section 4.1.5
 - o "Autonomous filters" typo (instead of Autonomous systems) corrected in the former section 4.2
 - o Removal of an example for manual address validation in former section 4.2.2.1
 - o RFC5735 obsoletes RFC3300
 - o Ingress/Egress replaced by Inbound/Outbound in all the document
- 12.2. 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:

- o 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
- o Change MD5 section with global TCP security session and introducing TCP-AO in former section 3.1. Added reference to BCP38
- o Added new section 3 about BGP router protection with forwarding plane ACL
- o Change text about prefix acceptable specificity in former section 4.1.3 to explain this doc does not try to make recommendations

- o 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
 - o Abstract reworded
 - o 6to4 exception described (only more specifics MUST be filtered)
 - o More specific -> more specifics
 - o should -> MUST for the prefixes an ISP needs to filter from its customers in former section 4.2.2.1
 - o Added "plus some headroom to permit growth" in former section 7
 - o Added new section on Next-Hop filtering
- 12.3. Diffs between draft-ietf-opsec-bgp-security-00 and draft-jdurand-bgp-security-02

Following changes have been made since previous document draft-jdurand-bgp-security-02:

- o Added a subsection for RTBH in next-hop section with reference to RFC6666
 - o Changed last sentence of introduction
 - o Many edits throughout the document
 - o Added definition of tier 1 transit provider
 - o Removed definition of a BGP peering
 - o Removed description of routing policies for IPv6 prefixes in IANA special registry as this now contains a routing scope field
 - o Added reference to RFC6598 and changed the IPv4 prefixes to be filtered by definition section
 - o IXP added in acronym/definition section and only term used throughout the doc now
- 12.4. Diffs between draft-ietf-opsec-bgp-security-01 and draft-ietf-opsec-bgp-security-00

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

- o Obsolete RFC2385 moved from normative to informative reference
- o Clarification of preference of TCP-AO over MD5 in former section 4.1
- o Mentioning KARP efforts in TCP session protection section in former section 4 and adding 3 RFC as informative references: 6518, 6862 and 6952
- o Removing reference to SIDR working-group
- o Better dissociating origin validation and path validation to clarify what's potentially available for deployment
- o Adding that SIDR mechanisms should be implemented in addition to the other ones mentioned throughout this document
- o Added a paragraph in former section 8 about ASN renumbering
- o Change of security considerations section
- o Added the newly created IANA IPv4 Special Purpose Address Registry instead of references to RFCs listing these addresses

12.5. Diffs between draft-ietf-opsec-bgp-security-02 and draft-ietf-opsec-bgp-security-01

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

- o Added a reference to draft-ietf-sidr-origin-ops
- o Added a reference to RFC6811 and RFC6907
- o Changes "Most of RIR's" to "A majority of RIR's" on IRR availability
- o Various edits
- o Added NIST BGP security recommendations document
- o Added that it's possible to get info from ISPs from RADB
- o Correction of the url for IPv4 special use prefixes repository
- o Clarification of the fact only prefixes with Global Scope set to False MUST be discarded

- o IANA list could be pulled directly by routers (not just pushed on routers).
 - o Warning added when prefixes are checked against IRR
 - o Recommend network operators to sign their routing objects
 - o Recommend network operators to publish their routing objects in IRR of their IRR when available
 - o Dissociate rules for local AS and downstreams in former section 5.1.4
- 12.6. Diffs between draft-ietf-opsec-bgp-security-03 and draft-ietf-opsec-bgp-security-02

Following changes have been made since previous document draft-ietf-opsec-bgp-security-02:

- o Added a note on TCP-AO to be preferred over MD5
- o Mention that loose AS filtering with customers can be combined with precise filters for important ASNs (example those of transits) that are must not be received on theses peers in former section 8.
- o MD5 removed from abstract
- o recommended -> RECOMMENDED where appropriate
- o Reference to BCP38 and BCP84 in former section 4.1
- o Added a note to RFC Editor to remove change section before publication
- o Removal of "future work" section
- o Added rate-limiting in addition to filtering in former section 3
- o Reference to IRRToolSet in former section 5.1.2.3
- o Removed "foreword" section

12.7. Diffs between draft-ietf-opsec-bgp-security-04 and draft-ietf-opsec-bgp-security-03

Following changes have been made since previous document draft-ietf-opsec-bgp-security-03:

- o RFC6890 updates RFC5735
- o RFC6890 updates RFC5156
- o Removed reference RFC2234 and RFC 4234
- o Moved route-server draft into informative reference section

12.8. Diffs between draft-ietf-opsec-bgp-security-05 and draft-ietf-opsec-bgp-security-04

Following changes have been made since previous document draft-ietf-opsec-bgp-security-04:

- o RFC7196 updates draft-ietf-idr-rfd-usable
- o RFC7115 updates draft-ietf-sidr-origin-ops
- o draft-ietf-idr-ix-bgp-route-server-05 updates ietf-idr-ix-bgp-route-server-00

12.9. Diffs between draft-ietf-opsec-bgp-security-06 and draft-ietf-opsec-bgp-security-05

Following changes have been made since previous document draft-ietf-opsec-bgp-security-05:

- o Wording improvements
- o Introduction improved
- o References are expanded (not just reference numbers are displayed but also the title of the document)
- o First occurrence of accronyms expanded
- o GTSM for multi-hop peerings
- o Remove eBGP as protected by BCP38
- o Add a caveat for IPsec for session protection

- o Changed MUST for SHOULD everywhere
- o Small changes in communities section
- o Removed simplified IPv6 prefix list
- o Removed note in section 9 about 32 bits ASN
- o IXP LAN prefix example in appendix
- o Make sure all references are in the text. Most of them were removed as they were initially here for previous version when IANA registries with routing scopes did not exist

13. Acknowledgements

The authors would like to thank the following people for their comments and support: Marc Blanchet, Ron Bonica, Randy Bush, David Freedman, Wesley George, Daniel Ginsburg, David Groves, Mike Hugues, Joel Jaeggli, Tim Kleefass, Warren Kumari, Jacques Latour, Lionel Morand, Jerome Nicolle, Hagen Paul Pfeifer, Thomas Pinaud, Carlos Pignataro, Jean Rebiffe, Donald Smith, Kotikalapudi Sriram, Matjaz Straus, Tony Tauber, Gunter Van de Velde, Sebastian Wiesinger, Matsuzaki Yoshinobu.

Authors would like to thank once again Gunter Van de Velde for presenting the draft at several IETF meetings in various working groups, indeed helping dissemination of this document and gathering of precious feedback.

14. IANA Considerations

This memo includes no request to IANA.

15. Security Considerations

This document is entirely about BGP operational security. It depicts best practices that one should adopt to secure its BGP infrastructure: protecting BGP router and BGP sessions, adopting consistent BGP prefix and AS-path filters and configure other options to secure the BGP network.

On the other hand this document doesn't aim at depicting existing BGP implementations and their potential vulnerabilities and ways they handle errors. It does not detail how protection could be enforced against attack techniques using crafted packets.

16. References

16.1. 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] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [3] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.
- [4] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.
- [5] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, January 2013.
- [6] Pelsser, C., Bush, R., Patel, K., Mohapatra, P., and O. Maennel, "Making Route Flap Damping Usable", RFC 7196, May 2014.

16.2. Informative References

- [7] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.
- [8] 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.
- [9] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.
- [10] Blunk, L., Damas, J., Parent, F., and A. Robachevsky, "Routing Policy Specification Language next generation (RPSLng)", RFC 4012, March 2005.
- [11] Dugal, D., Pignataro, C., and R. Dunn, "Protecting the Router Control Plane", RFC 6192, March 2011.
- [12] Lepinski, M. and S. Kent, "An Infrastructure to Support Secure Internet Routing", RFC 6480, February 2012.

- [13] Hilliard, N. and D. Freedman, "A Discard Prefix for IPv6", RFC 6666, August 2012.
- [14] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, May 2013.
- [15] Bush, R., "Origin Validation Operation Based on the Resource Public Key Infrastructure (RPKI)", BCP 185, RFC 7115, January 2014.
- [16] Kent, S. and A. Chi, "Threat Model for BGP Path Security", RFC 7132, February 2014.
- [17] "Internet Exchange Route Server",
<<http://tools.ietf.org/id/draft-ietf-idr-ix-bgp-route-server-05.txt>>.
- [18] Smith, P. and C. Panigl, "RIPE-378 - RIPE Routing Working Group Recommendations On Route-flap Damping", May 2006.
- [19] Smith, P., Evans, R., and M. Hughes, "RIPE-399 - RIPE Routing Working Group Recommendations on Route Aggregation", December 2006.
- [20] Smith, P. and R. Evans, "RIPE-532 - RIPE Routing Working Group Recommendations on IPv6 Route Aggregation", November 2011.
- [21] Smith, P., Bush, R., Kuhne, M., Pelsser, C., Maennel, O., Patel, K., Mohapatra, P., and R. Evans, "RIPE-580 - RIPE Routing Working Group Recommendations On Route-flap Damping", January 2013.
- [22] "IANA IPv4 Special Purpose Address Registry",
<<http://www.iana.org/assignments/iana-ipv4-special-registry/iana-ipv4-special-registry.xhtml>>.
- [23] "IANA IPv6 Special Purpose Address Registry",
<<http://www.iana.org/assignments/iana-ipv6-special-registry/iana-ipv6-special-registry.xml>>.
- [24] "IANA IPv4 Address Space Registry",
<<http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml>>.

- [25] "IANA IPv6 Address Space Registry",
<<http://www.iana.org/assignments/ipv6-unicast-address-assignments/ipv6-unicast-address-assignments.xml>>.
- [26] "Routing Assets Database", <<http://www.radb.net>>.
- [27] "Security Requirements for BGP Path Validation",
<<http://datatracker.ietf.org/doc/draft-ietf-sidr-bgpsec-reqs/>>.
- [28] "Autonomous System (AS) Migration Features and Their Effects on the BGP AS_PATH Attribute",
<<http://datatracker.ietf.org/doc/draft-ga-idr-as-migration/>>.
- [29] "IRRToolSet project page", <<http://irrtoolset.isc.org>>.
- [30] Cooper, D., Heilman, E., Brogle, K., Reyzin, L., and S. Goldberg, "On the Risk of Misbehaving RPKI Authorities",
<<http://www.cs.bu.edu/~goldbe/papers/hotRPKI.pdf>>.

Appendix A. IXP LAN prefix filtering - example

An IXP in the RIPE region is allocated an IPv4 /22 prefix by RIPE NCC (X.Y.0.0/22 in this example) and uses 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 SHOULD 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 an eBGP peering (most likely from 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 6.1.2.2.1)

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.

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@ipspace.net

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

Email: gert@space.net