NOCTION
NETWORK INTELLIGENCE

BGP Filtering Best Practices

www.noction.com

Table of Contents

Introduction

BGP filtering is used to control prefixes that are received and advertised to BGP peers. Filtering is critically important at Tier1, Tier2 and Tier3 levels as it can restrain and eliminate the damage to your network and from your network. However, a large number of routine BGP leaks and hijacks is downright evidence of the poor use of filters at all levels.

Normally, network operators should have filters in place to make sure they only accept the correct prefixes from their customers. Prefixes exchanged between BGP peers should be controlled with inbound and outbound filters that can match on IP prefixes, AS paths or any other attributes of a BGP prefix (e.g. BGP communities).

The Internet Routing Registry (IRR) is a database containing Internet routing information used by network operators to register their assigned network resources. It is possible to use the IRR information to build a list of originated or transited prefixes that one may accept for a given neighbor AS. This can be easily done, using scripts as well as existing tools capable of retrieving information from the registries. Prefixes from the entire AS can be obtained by entering the 'whois' commands. For instance, to list all prefixes originated on AS1759 against the Routing Assets Database (RADb), issue the command below.

```
$ whois -h whois.radb.net -- '-i origin AS1759' |
grep -Eo "([0-9.]+){4}/[0-9]+"
```

```
debian@debian:~$
debian@debian:~$ whois -h whois.radb.net -- '-i origin AS1759' | grep -Eo "([0-9.]+){4}/[0-9]+"
139.157.0.0/16
147.44.0.0/16
193.178.133.0/24
149.197.0.0/16
144.4.0.0/16
93.106.0.0/16
192.83.96.0/22
192.58.44.0/22
131.177.0.0/16
193.142.8.0/22
193.142.13.0/24
193.142.14.0/23
212.182.192.0/18
```

Picture 1 - *Prefixes Announced by AS1759* **<Output is truncated>**

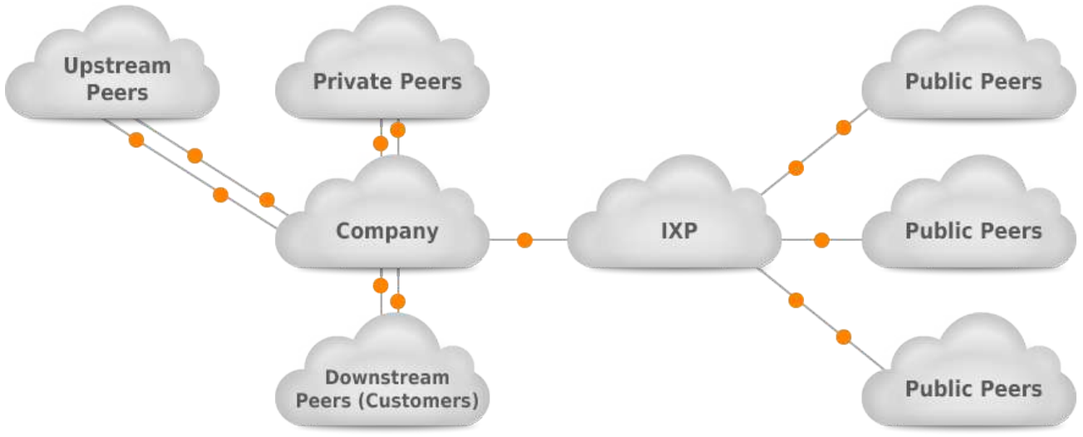
NOTE: Projects such as IRRToolSet provide tools that can be used to simplify the creation of an automated filter configuration from policies stored in IRR.

The RIR database cannot be completely trusted as some database maintainers do not check whether an entry is made by a real owner or if prefixes were transferred to someone else.

1. Different Peering Relationships

Let's take a look at different types of peering sessions at a company that is a Tier 2 provider depicted in Picture 2:

- **Public Peering**
- **Private (Bilateral) Peering**
- **Upstream (Transit Peering)**
- **Downstream (Customer Peering)**



Picture 2 - *Different Peering Sessions*

NOTE: Tier 1 providers do not buy transit service, they only peer to each other to maintain global routing reachability. When traffic is transferred free of charge from one network to the other, the relationship is called "settlement-free peering".

1.1 Public Peering

Public peering is a type of relationship where two ISPs exchange prefixes and traffic via a single public Internet Exchange Peering Point (IXP). Typically, each ISP participating in IXP brings its own border router that connects with the Ethernet port to the IXP LAN. The WAN port is used to connect back to the ISP network. The ISP's routers peer with each other (if no route-server is used) using eBGP and can freely exchange own prefixes and IP traffic (without fees). The benefit of public peering at IXP is that we can peer with many other ISPs over the IXP Layer2 device (switch). Therefore, we do not need to build a separate physical medium to connect to peers as it is done in the case of private peering.

NOTE: Internet Exchange Point network is a Layer 2 LAN.

1.2 Private or Bilateral Peering

In this relationship, two ISPs peer over a direct (private) transport link (e.g. 10Gb Ethernet fiber), typically without paying for exchanged traffic. Peers do not advertise a full routing table, but only prefixes originated by a peer as well as customers' prefixes. This type of peering makes sense for a larger volume of traffic between the two networks since it is cost-effective.

1.3 Upstream or Transit Peering

Unless we are a Tier1 provider, we buy transit service from an upstream (transit) provider that provides a full IPv4 and IPv6 routing tables for us or a default route. This allows us to reach any IPv4 or IPv6 host on the Internet. We, in our turn, announce locally originated prefixes to the upstream provider via eBGP, including the prefixes of our customers (Picture 2).

1.4 Downstream or Customer Peering

In customer peering, an ISP represents a transit provider for its customers. An ISP advertises a full Internet routing table to customers or a default route. The prefixes received from customers are advertised to other customers, transit providers, private and public peers.

2. Filters with Public/Private Peers

There are basically two methods for setting filters with peers. The first one is the loose method where no checks are performed against the RIR databases. When strict method is used, the announcements are strictly verified to conform to what is declared in the routing registries.

2.1 Inbound Filtering Loose Option

According to RFC7454, loose inbound filters for the received prefixes from a BGP peer should consider the following:

- Prefixes that are not globally routable [1]
- Prefixes not allocated by IANA - IPv6 only, IPv4 address space is depleted
- Routes that are too specific - IPv4 prefixes longer than /24 and IPv6 prefixes longer than /48
- Prefixes belonging to the local AS - filter own prefixes on peerings with all your peers
- IXP LAN prefixes - do not accept the more specific match for IXP LAN subnet from eBGP peers
- Default route - IPv4 0.0.0.0/0 or IPv6 ::/0 prefixes

2.2 Inbound Filtering Strict Option

Strict inbound filters are applied to make sure advertisements strictly conform to what is declared in the routing registries. In addition to

this, network administrators can apply the same filters as in the case of loose option when the routing registry that's used as the source of information by the script is not fully trusted.

2.3 Outbound Filtering

The configuration should ensure that only locally originated and customers' prefixes are sent. This can be achieved by using BGP communities, AS paths, or both. Moreover, it may be desirable to add the following filters before any policy to avoid the unwanted route announcements due to a bad configuration:

- **Prefixes that are not globally routable**
- **Routes that are too specific**
- **IXP LAN prefixes**
- **The default route**

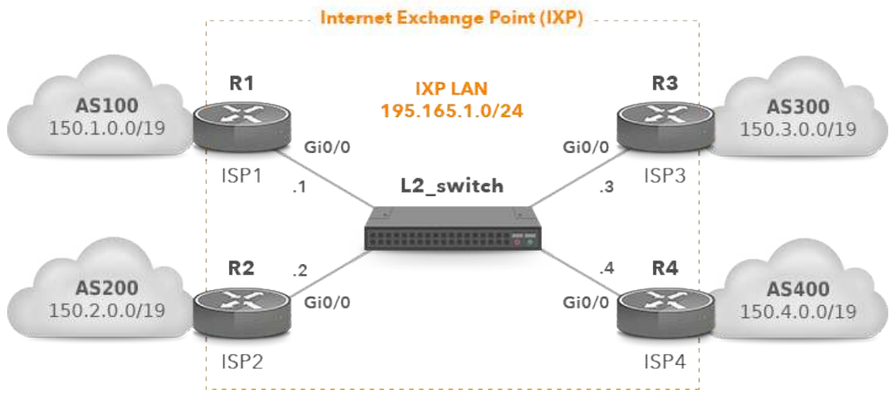
2.4 Strict Prefixes Filtering at IXP

The upcoming parts 2.4.1, 2.4.2 and 2.4.3 discuss different outbound filtering methods that we will configure for the IXP peer1 (AS100). As a result, ISP1's border router in AS100 will announce only own prefixes and customers' prefixes to IXP peers in AS200, AS300 and AS400. No default route or a full Internet routing table are announced to other IXP peers (or private peers). Carrying the default route or full routing table would create a possibility that IXP router and bandwidth will be stolen by non-peering members.

The first method is based on a prefix-list and the second one is using AS path filters-lists. The third method employs BGP communities. In all three cases, strict inbound prefix filtering is used to accept prefixes from other ISPs only if their peers are entitled to originate them.

NOTE: Since the configuration for other IXP routers follows the same concept, we will not discuss it.

Public IXP with four IXP participants is depicted in Picture 3. A managed switch is provided by the IXP operator and connects the ISPs' border routers (IXP's members). The switch is an L2 device and does not run BGP. The ISP border routers are configured to establish an eBGP relationship with each other over a shared IXP LAN subnet; each border router is configured for a different public AS number.



Picture 3 - *IXP with Four IXP Members Connected to L2 Switch*

The interface Gi0/0 is connected to the LAN IXP network and configured with a public IP address from the IXP LAN subnet.

```
interface GigabitEthernet0/0
 description LAN IXP
 ip address 195.165.1.1 255.255.255.0
```

NOTE: Some IXPs use private addresses for the IXP LAN subnet. Therefore, the IXP network cannot be leaked to the Internet as most ISPs filter the RFC1918 address space.

2.4.1 Filtering Outgoing BGP Updates with Prefix Lists

Firstly, we will create the peer-group ixp-peer so we can easily assign the same prefix-list pl-peer100 in outbound direction to the members of the peer-group.

```
router bgp 100
 neighbor ixp-peer peer-group
 neighbor ixp-peer send-community
 neighbor ixp-peer prefix-list pl-peer100 out
 neighbor ixp-peer route-map set-loc-pref-in in
```

The prefix-list pl-peer100 is matching the address block 150.1.0.0/19 assigned to AS100 that we will announce to the other IXP members.

```
ip prefix-list pl-peer100 seq 5 permit 150.1.0.0/19
```

Next, we will define IXP peers and assign them to an ixp-peer group. A different prefix-list configured for each IXP peer and applied inbound is matching a prefix received from the peer. As a result, AS100 will accept only prefixes assigned to other IXP peers; prefixes that are not matched by the prefix-lists are filtered. These strict inbound filters make sure advertisements strictly conform to what is declared in the routing registries.

```
router bgp 100
 neighbor 195.165.1.2 remote-as 200
 neighbor 195.165.1.2 peer-group ixp-peer
 neighbor 195.165.1.2 prefix-list pl-peer200 in
 neighbor 195.165.1.3 remote-as 300
 neighbor 195.165.1.3 peer-group ixp-peer
 neighbor 195.165.1.3 prefix-list pl-peer300 in
 neighbor 195.165.1.4 remote-as 400
 neighbor 195.165.1.4 peer-group ixp-peer
 neighbor 195.165.1.4 prefix-list pl-peer400 in

ip prefix-list pl-peer200 seq 5 permit 150.2.0.0/19
ip prefix-list pl-peer300 seq 5 permit 150.3.0.0/19
ip prefix-list pl-peer400 seq 5 permit 150.4.0.0/19
```

NOTE: The full Internet routing table is learned via ISP1's upstream provider. In order to prefer a path to the prefixes originated by IXP peers to the path learned from the ISP1's transit provider, we will configure a route-map set-loc-pref-in. The route-map applied inbound on ISP1's border router for the peer-group ixp-peer increases the local preference to 150 for the routes originated from IXP peers.

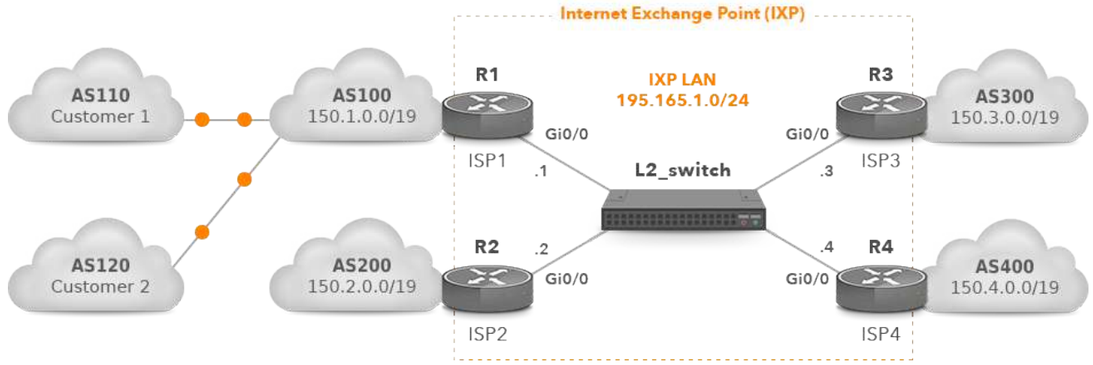
```
route-map set-loc-pref-in permit 10
 set local-preference 150
```

2.4.2 Filtering Outgoing BGP Updates with AS Path Filters

AS100 needs to know all the prefixes that its IXP peers are announcing to set inbound filters. Each time IXP peer announces a new prefix or a new customer prefix is added, prefix-lists must be updated in order to reflect this change.

The use of filter-lists for filtering outgoing routes to IXP peers represents an alternative and a more scalable approach to this issue. A filter-list configured for all IXP peers for outbound routes has to match local AS (AS100) and all downstream ASes assigned to AS100 customers.

Let's say that ISP1 (AS100) has two customers of its own - Customer1 (AS110) and Customer2 (AS120) (Picture 4). Only the local AS100 and the customers' AS are permitted by the filter list 10 for outgoing routes. For routes incoming from IXP peers, strict inbound filtering is used to ensure that IXP peers are entitled to announce the appropriate routes.



Picture 4 - *IXP with 4 IXP Participants*

```
router bgp 100
neighbor ixp-peer peer-group
neighbor ixp-peer send-community
neighbor ixp-peer prefix-list pl-bogons out
neighbor ixp-peer filter-list 10 out
neighbor ixp-peer route-map set-loc-pref-in in
neighbor 195.165.1.2 remote-as 200
neighbor 195.165.1.2 peer-group ixp-peer
neighbor 195.165.1.2 prefix-list pl-peer200 in
neighbor 195.165.1.3 remote-as 300
neighbor 195.165.1.3 peer-group ixp-peer
neighbor 195.165.1.3 prefix-list pl-peer300 in
```

```
neighbor 195.165.1.4 remote-as 400
neighbor 195.165.1.4 peer-group ixp-peer
neighbor 195.165.1.4 prefix-list pl-peer400 in
```

Now, we can define AS path access-list 10. The regular expression ^$ in ACL statement is matching the empty AS_PATH so it permits locally announced prefixes by AS100. The next two statements permit only locally announced prefixes from the AS110 and AS120.

```
ip as-path access-list 10 permit ^$
ip as-path access-list 10 permit ^110$
ip as-path access-list 10 permit ^120$
```

We need to configure prefix-lists matching routes that IXP peers are entitled to announce.

```
ip prefix-list pl-peer200 seq 5 permit 150.2.0.0/19
ip prefix-list pl-peer300 seq 5 permit 150.3.0.0/19
ip prefix-list pl-peer400 seq 5 permit 150.4.0.0/19
```

The AS path access-list 10 applied for the peer-group ixp-peer for outgoing routes is matching only the locally announced prefixes by AS100, AS110, and AS120. However, the policy is not perfect yet and prone to configuration errors as it still permits routes that should not be announced to IXP peers. Therefore, we need to employ the prefix-list pl-bogons applied to the peer-group ixp-peer for routes announced by AS100 to IXP peers.

The prefix-list is matching routes that are not globally routable, too specific routes, IXP LAN prefix, and the default route.

Bogons are prefixes that have not yet been allocated (IPv6 only) or are used for a special purpose. They are denied by sequence 5 - 45. The sequence 50 blocks announcing of IXP LAN network and the sequence 55 denies the default route. All other prefixes with the mask length /24 and less are permitted by sequence 100. The prefixes that are too specific (with the mask length from /25 to /32) won't be advertised to IXP peers.

```
ip prefix-list pl-bogons seq 5 deny 0.0.0.0/8 le 32
ip prefix-list pl-bogons seq 10 deny 10.0.0.0/8 le 32
ip prefix-list pl-bogons seq 15 deny 127.0.0.0/8 le 32
ip prefix-list pl-bogons seq 20 deny 169.254.0.0/16 le 32
ip prefix-list pl-bogons seq 25 deny 172.16.0.0/12 le 32
ip prefix-list pl-bogons seq 30 deny 192.0.2.0/24 le 32
ip prefix-list pl-bogons seq 35 deny 192.168.0.0/16 le 32
ip prefix-list pl-bogons seq 40 deny 224.0.0.0/4 le 32
ip prefix-list pl-bogons seq 45 deny 240.0.0.0/4 le 32
ip prefix-list pl-bogons seq 55 deny 0.0.0.0/0
ip prefix-list pl-bogons seq 100 permit 0.0.0.0/0 le 24
```

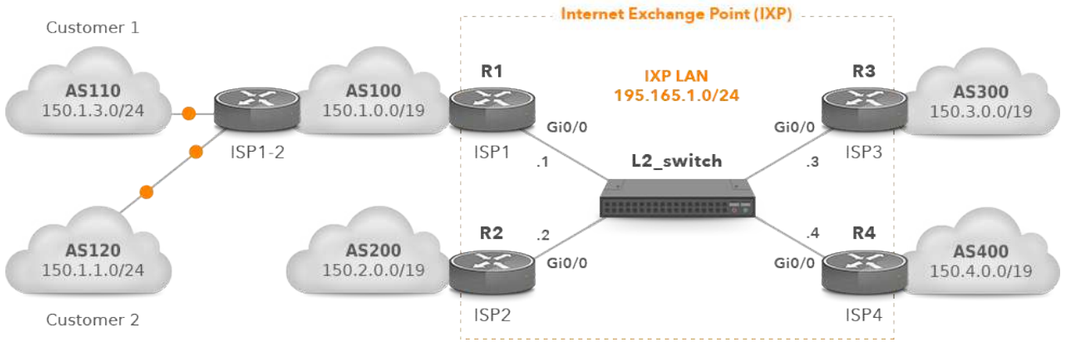
Prefixes learned from IXP peers will have a higher preference (150) than those learned from the upstream ISP.

```
route-map set-loc-pref-in permit 10
 set local-preference 150
```

2.4.3 Filtering Outgoing BGP Updates with BGP Communities

The AS path access-list must be updated when a new customer is added. Our configuration will not scale as more and more BGP customers are added to AS100. The scalable solution relies on BGP communities to determine what is announced to IXP peers. Thanks to the use of communities, we do not need to update any as-path filters, prefix-lists when a BGP customer announces more prefixes. Only the filters at the aggregation edge need to be updated and those new prefixes will automatically be tagged with the community to allow them through to AS100's IXP peers.

The AS130 and AS110 are now customers of ISP1 (AS100) and they are single-homed to the customer aggregation router ISP1-2 (AS100) (Picture 5). ISP1 provides transit service to them using both upstream provider connection (not shown) and public peering at IXP. Routers ISP1-2 and ISP1 peer each other using iBGP.



Picture 5 - *AS100 Providing Transit Service for AS130 and AS110*

Router ISP1-2 (Customer Aggregation Router)

The route-map comm is applied to the aggregated route 150.1.0.0/19 in order to set the BGP community 100:1000. The route-map cust-pol-in is applied to incoming routes from both customers to set the community 100:1100. Router ISP1-2 set itself as a BGP next-hop for routes received from IXP peers.

```
router bgp 100
 network 150.1.0.0 mask 255.255.224.0 route-map comm
 neighbor 150.1.5.2 remote-as 110
 neighbor 150.1.5.2 prefix-list pl-as110-in in
 neighbor 150.1.5.2 prefix-list default out
 neighbor 150.1.5.2 route-map cust-pol-in in
 neighbor 150.1.5.5 remote-as 100
 neighbor 150.1.5.5 next-hop-self
 neighbor 150.1.5.5 send-community
 neighbor 150.1.5.10 remote-as 130
 neighbor 150.1.5.10 prefix-list pl-as130-in in
 neighbor 150.1.5.10 prefix-list default out
 neighbor 150.1.5.10 route-map cust-pol-in in
```

Configure a new BGP community format aa:nn.

```
ip bgp-community new-format
```

The prefix-list default is matching a default route. Implicit deny at the end of the prefix-list denies other prefixes.

```
ip prefix-list default seq 5 permit 0.0.0.0/0
```

We need to configure prefix-lists for each customer, matching the routes that are customers entitled to announce. AS100 needs to update prefix-lists every time customers add new prefixes.

```
ip prefix-list pl-as110-in seq 5 permit 150.1.1.0/24
ip prefix-list pl-as130-in seq 5 permit 150.1.3.0/24
```

BGP community 100:1000 is attached to the aggregated route 150.1.0.0/19 on the customers' aggregation router ISP1-2. BGP community 100:1100 is attached to all BGP customer prefixes.

```
route-map comm permit 10
 set community 100:1000

route-map cust-pol-in permit 10
 set community 100:1100
```

Router ISP1 (IXP Peer)

The route-map ixp-peers-out is applied to outbound routes for the peer-group ixp-peer (eBGP peers IXP2, IXP3, and IXP4). The route-map is matching community lists, Therefore, IXP1 advertises only the routes with the attached communities 100:1000 and 100:1100 to other IXP peers.

The strict inbound policy is unique for each IXP peer and is based on prefix-lists. Peers need to update ISP if filters need to be changed.

```
router bgp 100
 neighbor ixp-peer peer-group
 neighbor ixp-peer send-community
 neighbor ixp-peer route-map set-local-pref in
 neighbor ixp-peer route-map ixp-peers-out out
 neighbor 150.1.5.6 remote-as 100
 neighbor 195.165.1.2 remote-as 200
 neighbor 195.165.1.2 peer-group ixp-peer
 neighbor 195.165.1.2 prefix-list pl-peer200 in
 neighbor 195.165.1.3 remote-as 300
 neighbor 195.165.1.3 peer-group ixp-peer
 neighbor 195.165.1.3 prefix-list pl-peer300 in
 neighbor 195.165.1.4 remote-as 400
 neighbor 195.165.1.4 peer-group ixp-peer
 neighbor 195.165.1.4 prefix-list pl-peer400 in

ip bgp-community new-format
```

Define community list for the aggregated route and customers routes.

```
ip community-list 10 permit 100:1000
ip community-list 11 permit 100:1100
```

We must define prefix-list for each IXP peer.

```
ip prefix-list pl-peer200 seq 5 permit 150.2.0.0/19
ip prefix-list pl-peer300 seq 5 permit 150.3.0.0/19
ip prefix-list pl-peer400 seq 5 permit 150.4.0.0/19
```

The route-map ixp-peers-out is matching community lists 10 and 11. Hidden implicit deny at the end of the route-map ensures that routes without an attached community are filtered.

```
route-map ixp-peers-out permit 10
 match community 10 11
```

Again, the routes received from IXP peers have higher local preference.

```
route-map set-local-pref permit 10
 set local-preference 150
```

Router IXP1 advertises the aggregate route 150.1.0.0/19 to all IXP peers. BGP table of IXP2 with this route and attached community 100:1000 is depicted on Picture 6.

```
IXP2#
IXP2#show bgp 150.1.0.0/19
BGP routing table entry for 150.1.0.0/19, version 42
Paths: (1 available, best #1, table default)
  Not advertised to any peer
  Refresh Epoch 6
  100
    195.165.1.1 from 195.165.1.1 (195.165.1.1)
      Origin IGP, localpref 100, valid, external, best
      Community: 100:1000
      rx pathid: 0, tx pathid: 0x0
IXP2#
```

Picture 6 - *Aggregated Route 150.1.0.0/19 Presented in the IXP2 BGP Table*

The customers' routes are received with the community 100:1100 (Picture 7).

```
IXP2#
IXP2#show bgp 150.1.1.0
BGP routing table entry for 150.1.1.0/24, version 40
Paths: (1 available, best #1, table default)
  Not advertised to any peer
  Refresh Epoch 6
  100 110
    195.165.1.1 from 195.165.1.1 (195.165.1.1)
      Origin IGP, localpref 100, valid, external, best
      Community: 100:1100
      rx pathid: 0, tx pathid: 0x0
IXP2#
```

Picture 7 - *Customer Route 150.1.1.0/24 Presented in IXP2 BGP Table*

2.5 Redistributing IXP LAN Prefixes to iBGP Peers

ISP border routers at IXP should not be configured to carry the IXP LAN prefixes within IGP or iBGP. IXP router interfaces are often the target of DDoS attacks. In such attacks, IXP needs to be able to stop announcing the IXP prefix. If we redistribute the peering LAN into OSPF or ISIS networks, we cannot control the propagation of the prefix in our network. Therefore, configure next-hop-self for all iBGP peers.

Picture 8 depicts a BGP table of the router ISP1-2 (AS100) with ISP2 prefix 150.2.0.0/10 and the next-hop IP 195.165.1.2 (IXP2 LAN interface). Router ISP1-2 is the iBGP peer of the router IXP1 (AS100) (Picture 5).

```
ISP1-2#show bgp 150.2.0.0
BGP routing table entry for 150.2.0.0/19, version 8
Paths: (1 available, best #1, table default)
  Not advertised to any peer
  Refresh Epoch 1
  200
    195.165.1.2 from 150.1.5.5 (195.165.1.1)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0
ISP1-2#
```

Picture 8 - *ISP1-2 BGP Table with IXP2 Prefix 150.2.0.0/19 and Next-hop 195.165.1.2 IXP2 Interface) Before Applying Next-hop-self to iBGP peers on IXP1*

```
router bgp 100
 neighbor ibgp-peer peer-group
 neighbor ibgp-peer next-hop-self
 neighbor 150.1.5.6 remote-as 100
 neighbor 150.1.5.6 peer-group ibgp-peers
```

When router IXP1 (AS100) advertises a prefix received from IXP peers via iBGP to the iBGP peer 150.1.5.6, it will change the next-hop to itself, which the iBGP peer will have a route to. Picture 9 depicts a new next-hop address 150.1.5.5 for the ISP2 prefix 150.2.0.0/10 on the router ISP1-2 (AS100). The IP 150.1.5.5 is the address of iBGP neighbor – router IXP1.

```
ISP1-2#show bgp 150.2.0.0
BGP routing table entry for 150.2.0.0/19, version 10
Paths: (1 available, best #1, table default)
  Not advertised to any peer
  Refresh Epoch 2
  200
    150.1.5.5 from 150.1.5.5 (195.165.1.1)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0
ISP1-2#
```

Picture 9 - *ISP1-2 BGP Table with IXP2 Prefix 150.2.0.0/19 and Next-hop Set to 150.1.5.5 After Applying Next-hop-self to iBGP peers on IXP1*

2.6 BGP Next-hop Filtering for Public Peers

eBGP allows setting the BGP next-hop address to be any address on the IXP. As a result, traffic can be directed to a different destination than to the IXP peer announcing the prefix. This is the desired behavior when a route-server is used at IXP to scale full BGP mesh. All prefixes sent to a router server are distributed to all ASNs that peer with the server. Typically, BGP route-server has no capacity to receive traffic. Therefore, BGP route server will announce prefixes with a next-hop setting pointing to the router that originally announced the prefix to the route server. For this reason, we accept the next-hop attribute advertised by the route-server.

The situation is different when direct peering (without route-server) is used at an IXP. In such case, RFC7457 recommends applying the inbound BGP policy on IXP peering so the next-hop for an accepted prefix is set to BGP peer address (belonging to the IXP LAN) that sent the prefix. This prevents one peer from tricking the other one into sending packets into a black hole (unreachable next hop) or to an unsuspecting third party who would then have to carry the traffic.

Let's say that ISP2 (AS200) will announce its prefix 150.2.0.0/19 with the next-hop IP 195.165.1.3 (IXP3) to IXP1. As a result, ISP1 (AS100) will send traffic destined for 150.2.0.0/19 to the router IXP3 instead of IXP2 (Picture 10).

```
IXP1#show bgp 150.2.0.0
BGP routing table entry for 150.2.0.0/19, version 13
Paths: (1 available, best #1, table default)
  Advertised to update-groups:
     1
  Refresh Epoch 1
  200
    195.165.1.3 from 195.165.1.2 (195.165.1.2)
      Origin IGP, metric 0, localpref 150, valid, external, best
      rx pathid: 0, tx pathid: 0x0
IXP1#
```

Picture 10 - *ISP2's Prefix 150.2.0.0/19 with Incorrect Next-Hop 195.165.1.3 in BGP Table of IXP1*

Let's add the route-maps to the existing IXP1 configuration based on BGP communities that set correct next-hop for IXP peers IXP2, IXP3, and IXP4. The route-maps ixp-peer200-in, ixp-peer300-in and ixp-peer400-in are applied for incoming routes from IXP peer.

```
router bgp 100
 neighbor ibgp-peer peer-group
 neighbor ixp-peer peer-group
 neighbor ixp-peer send-community
 neighbor ixp-peer route-map set-local-pref in
 neighbor ixp-peer route-map ixp-peers-out out
 neighbor 150.1.5.6 remote-as 100
 neighbor 150.1.5.6 peer-group ibgp-peer
 neighbor 195.165.1.2 remote-as 200
 neighbor 195.165.1.2 peer-group ixp-peer
 neighbor 195.165.1.2 route-map ixp-peer200-in in
```

```
 neighbor 195.165.1.3 remote-as 300
 neighbor 195.165.1.3 peer-group ixp-peer
 neighbor 195.165.1.3 route-map ixp-peer300 in
 neighbor 195.165.1.4 remote-as 400
 neighbor 195.165.1.4 peer-group ixp-peer
 neighbor 195.165.1.4 route-map ixp-peer400 in

ip bgp-community new-format
```

Define community list for the aggregated route and customers routes.

```
ip community-list 10 permit 100:1000
ip community-list 11 permit 100:1100
```

We must define prefix-list for each IXP peer.

```
ip prefix-list pl-peer200 seq 5 permit 150.2.0.0/19
ip prefix-list pl-peer300 seq 5 permit 150.3.0.0/19
ip prefix-list pl-peer400 seq 5 permit 150.4.0.0/19
```

Routes received from all IXP peers have higher local preference.

```
route-map set-local-pref permit 10
 set local-preference 150
```

Each route-map matches the IXP peer's prefixes and sets the received next-hop to IXP peer IP address. Hidden implicit deny at the end of each route-map filters the received prefixes that are not defined in the prefix-list for each IXP peer.

```
route-map ixp-peer200-in permit 10
 match ip address prefix-list pl-peer200
 set ip next-hop 195.165.1.2

route-map ixp-peer300-in permit 10
 match ip address prefix-list pl-peer300
 set ip next-hop 195.165.1.3

route-map ixp-peer400-in permit 10
 match ip address prefix-list pl-peer400
 set ip next-hop 195.165.1.4
```

The route-map ixp-peers-out is matching community lists 10 and 11.

```
route-map ixp-peers-out permit 10
 match community 10 11
```

When the route-map ixp-peer200-in is applied for the IXP peer2 - 195.165.1.2, the next-hop attribute is changed 195.165.1.2 (IXP2) (Picture 11).

```
IXP1#show bgp 150.2.0.0
BGP routing table entry for 150.2.0.0/19, version 8
Paths: (1 available, best #1, table default)
  Advertised to update-groups:
     2
  Refresh Epoch 2
  200
    195.165.1.2 from 195.165.1.2 (195.165.1.2)
      Origin IGP, metric 0, localpref 100, valid, external, best
      rx pathid: 0, tx pathid: 0x0
IXP1#
```

Picture 11 - *ISP2's Prefix 150.2.0.0/19 with Correct Next-Hop 195.165.1.2 in BGP Table of IXP1*

NOTE: We can also configure a strict input policy that filters updates from IXP if correct next-hop IP is not present. Below is an example for IXP peer IXP2 configured on the router IXP1.

```
ip prefix-list pl-peer200-nh seq 5 permit 195.165.1.2/32

route-map ixp-peer200-nh-in permit 10
 match ip next-hop prefix-list pl-peer200-nh

router bgp 100
 neighbor 195.165.1.2 route-map ixp-peer200-nh-in in
```

2.7 Maximum Prefixes on Public/Private Peering

RFC7457 recommends limiting the number of routes received from public/private peers. The number for a peer should be set to a lower limit than the number of routes in the Internet. When a number of received routes from an IXP peer exceeds the value, ISP shutdowns an eBGP session with an IXP peer. This helps to deal with a situation when a peer accidentally advertises a full routing table. A network administrator can configure a limit that is 2 times higher than the typical number of routes received from a peer.

The command maximum-prefix limits an accepted number of prefixes received from the peer 195.165.1.2 to 200. If the number is exceeded, BGP peering session will be disabled until it is cleared down with the command clear ip bgp 195.165.1.2.

```
router bgp 100
 neighbor 195.165.1.2 maximum-prefix 200
```

We can configure a device to automatically re-establish a BGP neighbor peering session when the peering session has been disabled. The time interval at which peering can be reestablished automatically is set to 30 minutes. If the number of routes exceeds the configured prefix limit, a peering session is brought down by a device.

```
router bgp 100
 neighbor 195.165.1.2 maximum-prefix 200 restart 30
```

The restart interval at which peering can be re-established is automatically set to 30 minutes.

3. Filters with Upstream (transit) Providers

3.1 Inbound Filtering

RFC7454 provides similar recommendations for route filtering from public/private peers and upstream providers if the full Internet routing table is desired from the upstream. The following routes should not be accepted from the peer (except the default route if it is required).

- **Special-Purpose Prefixes**
- **Unallocated Prefixes**
- **Prefixes that are too specific**
- **Prefixes belonging to the local AS**
- **Default Route - depending on whether or not the route is requested**

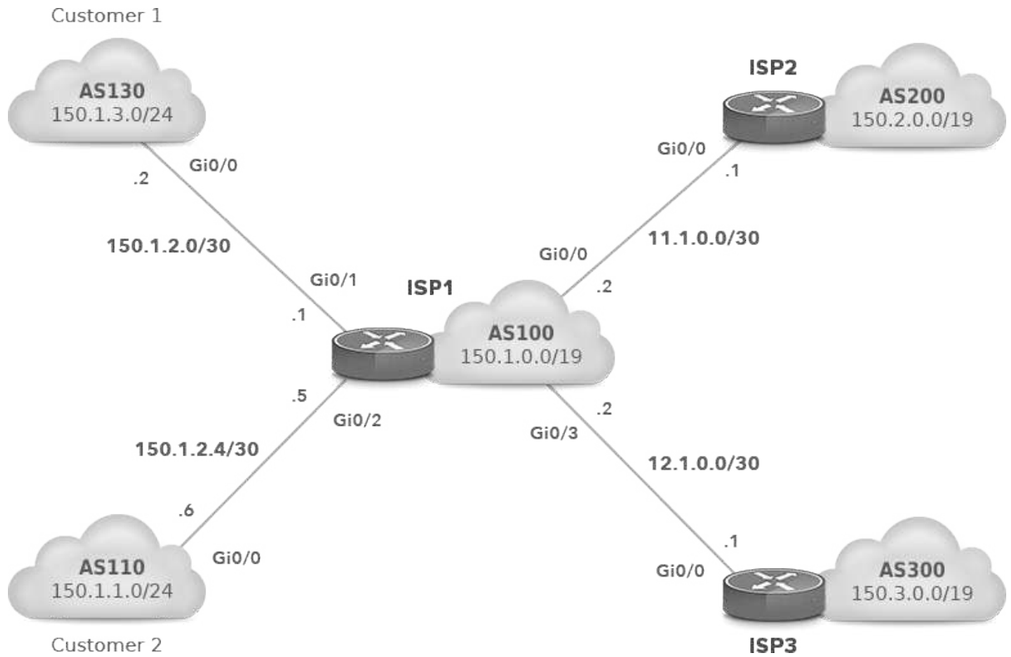
3.2 Outbound Filtering

We need to make sure that only authorized prefixes (those advertised by your AS and downstream customers) are being sent. In addition, the following prefixes should not be sent to peers.

- **Special-Purpose Prefixes**
- **Unallocated Prefixes**
- **Prefixes that are too specific**
- **Default Route - depending on whether or not the route is requested**

ISP1 (AS100) is multi-homed to ISP2 (AS200) and ISP3 (AS300). We will accept a full routing table from ISP2 and a default route from ISP3 (Picture 12). In order to conform to RFC7454, we will create inbound and outbound filters and apply them to the ISP1 router.

3.3 Filters with Upstream Providers Configuration



Picture 12 - *AS100 with two Upstream Providers ISP1 and ISP2*

```
router bgp 100
 network 150.1.0.0 mask 255.255.224.0
 neighbor upstream-group peer-group
 neighbor upstream-group prefix-list pl-upstream out
 neighbor upstream-group filter-list 10 out
 neighbor 11.1.0.1 remote-as 200
 neighbor 11.1.0.1 peer-group upstream-group
 neighbor 11.1.0.1 description ISP2
 neighbor 11.1.0.1 prefix-list pl-isp2 in
 neighbor 12.1.0.1 remote-as 300
 neighbor 12.1.0.1 peer-group upstream-group
 neighbor 12.1.0.1 description ISP3
 neighbor 12.1.0.1 prefix-list pl-isp3 in
 neighbor 150.1.2.2 remote-as 130
 neighbor 150.1.2.2 description Customer1
 neighbor 150.1.2.6 remote-as 110
 neighbor 150.1.2.6 description Customer2
```

The prefix-list pl-isp2 is applied to incoming routes from ISP2 (AS200). The prefix-list list does no accept special-purpose prefixes, local prefix 150.1.0.0/19 (and longer) and a default route. Only the routes that are longer than /25 are accepted.

```
ip prefix-list pl-isp2 deny 0.0.0.0/8 le 32
ip prefix-list pl-isp2 deny 10.0.0.0/8 le 32
ip prefix-list pl-isp2 deny 127.0.0.0/8 le 32
ip prefix-list pl-isp2 deny 169.254.0.0/16 le 32
ip prefix-list pl-isp2 deny 172.16.0.0/12 le 32
ip prefix-list pl-isp2 deny 192.0.2.0/24 le 32
ip prefix-list pl-isp2 deny 192.168.0.0/16 le 32
ip prefix-list pl-isp2 deny 224.0.0.0/3 le 32
ip prefix-list pl-isp2 deny 150.1.0.0/19 le 32
ip prefix-list pl-isp2 deny 0.0.0.0/0
ip prefix-list pl-isp2 permit 0.0.0.0/0 le 24
```

The prefix-list pl-isp3 is applied to incoming routes from ISP3 (AS300). The prefix-list is matching a default route, all other routes are filtered.

```
ip prefix-list pl-isp3 permit 0.0.0.0/0
```

The AS path access-list 10 is applied for outgoing routes from AS100. This prevents AS100 from becoming a transit provider for ISP1 and ISP2. The regular expression ^$ in the ACL statement is matching the empty AS_PATH so it permits locally announced prefixes by AS100. The next two statements permit only locally announced prefixes from AS110 and AS130.

```
ip as-path access-list 10 permit ^$
ip as-path access-list 10 permit ^110$
ip as-path access-list 10 permit ^130$
```

We still need to deny sending special-purpose prefixes, default route, and prefixes that are longer than /25 to both transit providers. The prefix-list pl-upstream applied to both ISPs filter these routes.

```
ip prefix-list pl-upstream deny 0.0.0.0/8 le 32
ip prefix-list pl-upstream deny 10.0.0.0/8 le 32
ip prefix-list pl-upstream deny 127.0.0.0/8 le 32
ip prefix-list pl-upstream deny 169.254.0.0/16 le 32
ip prefix-list pl-upstream deny 172.16.0.0/12 le 32
ip prefix-list pl-upstream deny 192.0.2.0/24 le 32
ip prefix-list pl-upstream deny 192.168.0.0/16 le 32
ip prefix-list pl-upstream deny 224.0.0.0/3 le 32
ip prefix-list pl-upstream deny 0.0.0.0/0
ip prefix-list pl-upstream permit 0.0.0.0/0 le 24
```

3.4 Maximum Prefixes on Upstream Providers

RFC7457 recommends limiting the number of routes received from upstream providers. From upstream providers that provide full routing, a limit should be higher than the number of routes on the Internet. However, a limit must be lower than the maximum number of entries in a TCAM memory. If not, a router may switch to slow software switching and crash due to higher CPU load.

When a number of received routes from an upstream provider exceeds the configured value, an ISP shutdowns the eBGP session with the transit provider.

4. Filters with Downstream (Customers)

4.1 Inbound Filtering

ISP should set a strict inbound policy that allows only prefixes assigned to customers. All other routes received from customers must be filtered. In case, a customer advertises too many prefixes, a loose policy can be applied, however, we still need to filter the following prefixes:

- **Special-Purpose Prefixes**
- **Unallocated Prefixes (IPv6 only)**
- **Prefixes that are too specific**
- **Prefixes belonging to the local AS**
- **Default Route**

4.2 Outbound Filtering

The configuration of outbound filters toward customers merely depends on their preferences. For instance, one customer prefers to receive a full routing table while another needs a default route only. Moreover, it can be a combination of some prefixes and a default route. According to RFC7457, if the customer wishes to receive a full routing table, the following prefixes should be filtered.

- **Prefixes that are not globally routable**
- **Routes that are too specific**
- **The default route**

4.3 Filters with Customers Configuration

We will use the same topology depicted in Picture 12. The prefix-lists pl-customer1-in and pl-customer-2-in are matching the prefixes assigned to ISP1 customers. The strict inbound policy configured on the aggregation router ISP1 ensures that only prefixes assigned to customers are allowed; all other prefixes are filtered.

```
router bgp 100
 neighbor 150.1.2.2 remote-as 130
 neighbor 150.1.2.2 description Customer1
 neighbor 150.1.2.2 prefix-list pl-as130-in in
 neighbor 150.1.2.2 prefix-list pl-as130-out out
 neighbor 150.1.2.6 remote-as 110
 neighbor 150.1.2.6 description Customer2
 neighbor 150.1.2.6 prefix-list pl-as110-in in
 neighbor 150.1.2.6 prefix-list pl-as110-out out
```

Define a prefix-list for each customer matching a prefix assigned to the customer.

```
ip prefix-list pl-as130-in permit 150.1.3.0/24
ip prefix-list pl-as110-in permit 150.1.1.0/24
```

Customer1 (AS130) wants to receive a default route only. Therefore, the prefix-list pl-as130-out is matching the default route.

```
ip prefix-list pl-as130-out permit 0.0.0.0/0
```

Customer2 (AS110) wants to receive the full Internet routing table. Following RFC7457, the prefix-list pl-as110-out filters prefixes that are not globally routable, too specific prefixes (with the length /25 and more) and the default route.

```
ip prefix-list pl-as110-out deny 0.0.0.0/8 le 32
ip prefix-list pl-as110-out deny 10.0.0.0/8 le 32
ip prefix-list pl-as110-out deny 127.0.0.0/8 le 32
ip prefix-list pl-as110-out deny 169.254.0.0/16 le 32
ip prefix-list pl-as110-out deny 172.16.0.0/12 le 32
ip prefix-list pl-as110-out deny 192.0.2.0/24 le 32
ip prefix-list pl-as110-out deny 192.168.0.0/16 le 32
ip prefix-list pl-as110-out deny 224.0.0.0/3 le 32
ip prefix-list pl-as110-out deny 0.0.0.0/0
ip prefix-list pl-as110-out permit 0.0.0.0/0 le 24
```

5. Protection Against BGP Leaks & Hijacks

5.1 BGP Leaks and Hijacks

BGP leaks are illegitimate advertisements that are caused by misconfiguration, mostly due to improper filtering. The AS that leaks the prefix inserts itself somewhere in the AS path. As a result, traffic is redirected through the leaker AS which can impact network performance. For instance, if ISP1 (AS100) leaks prefix 150.2.0.0/19 received from its upstream provider ISP1 to another upstream provider AS300, it may become a transit provider for the prefix (Picture 12). Packets sent from AS300 customers to the server with the IP address 150.2.0.1./24 (AS200) may follow the suboptimal path through AS100, however, they will finally reach the server.

Unlike BGP leaks which have no harmful purpose, prefix hijacking is planned and performed with a particular intention, e.g. conducting the man-in-the-middle or DoS attacks. The Autonomous System that hijacks a prefix claims that it is an originator of the prefix. In this case, the AS number of the hijacker is placed on the right side of the AS path. For instance, if AS100 starts announcing the subprefix 150.2.0.0/24 to both upstreams, the inbound policy configured on ISP2 (AS200) will very likely filter the subprefix as belonging to a local AS. However, if the loose inbound filters are configured on the router ISP3 (AS300), the router will install the route 150.2.0.0/24 into its BGP table. Remember, RFC7454 provides a recommendation about filtering too specific routes with the prefix length /25 and longer. Subsequently, the prefix 150.2.0.0/24 is being advertised to the other BGP peers and to the whole Internet, with the AS100 as an originator of the prefix. All other BGP speaking routers will send traffic destined for the server

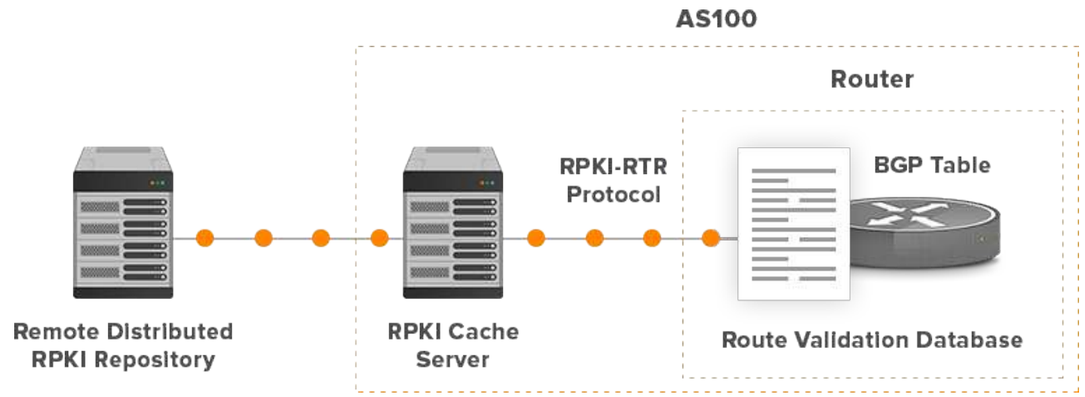
150.2.0.1/24 to AS100. Why? If there are prefixes 150.2.0.0/19 and 150.2.0.0/24 installed in the routing table of a router, the prefix 150.2.0.0/24 is the longest match and the more specific match always wins (regardless of AS_PATH or any BGP attribute).

Current BGP design is based on unconditional trust between BGP peers. BGP has no ability to verify the accuracy of routing information thus anyone can announce anything. Filters that we have already mentioned cannot fully protect us against BGP leaks or hijacks if they are not implemented globally on the Internet. We need to validate whether the AS number claiming to originate an address prefix (right-most AS number in the AS_PATH attribute) of the BGP route is authorized to announce the prefix. To do so, we need to enable Origin validation feature on a BGP speaking router and setup the Resource Public Key Infrastructure (RPKI) cache server.

5.2 Protection Against BGP Leaks

Origin validation is a mechanism defined in RFC6811 by which route advertisements can be authenticated as originating from an expected AS. To authenticate a prefix, the router (BGP speaker) queries the local database of validated prefix-to-AS mappings which are downloaded from the cache server. It is the job of the RPKI cache server to perform public-key validations, and generate a validated database of prefix-to-AS mappings. The router uses RPKI-Router protocol (RTR) to download a list of prefixes and permitted origin AS numbers from RPKI cache server. This approach does not require any change of BGP protocol; a router is completely unaware of RPKI infrastructure.

RPKI cache server maintains a local cache of the entire remote distributed repository collection by regularly synchronizing each element in the local cache against a repository (Picture 13). The repository contains Routing Origin Authorization (ROA) that are digital X.509 certificates, generated by operators of ASs and signed by their private key. ROAs contain a set of prefixes tied with the origin AS. The cache server, however, does not export ROAs to the router because ROAs are not directly used in route validation. Instead, the cache server generates a simplified version of the ROA to the router as an RV record (Juniper) or SOVC record (Cisco). RFC6811 refers to these objects as Validated ROA Payload (VRP).



Picture 13 - *RPKI Cache Server for AS100*

An RV record is a (prefix, maximum length, origin AS) triple. Overlapping prefixes are allowed. An example of the RV record is presented below. When the maximum length is not present, the AS is only authorized to advertise exactly the prefix specified in the RV record.

```
150.1.0.0/19-24 AS 100
150.1.1.0/24 AS 110
150.1.2.0/24 AS 120
150.1.3.0/24 AS 130
```

The router evaluates routes received from eBGP peer against the RV database. It matches any route received in the BGP UPDATE whose prefix matches the RV prefix, whose prefix length does not exceed the maximum length given in the RV record, and whose origin AS equals the origin AS given in the RV record. Such prefixes qualify as Valid routes. A prefix marked as Valid is installed in the BGP routing table.

***NOTE:** The origin of a route is represented by the right-most AS number in the AS_PATH attribute.*

A prefix is marked Invalid if it is found in RV database, but either the corresponding AS received from the eBGP peer is not the AS that appears in RV record, or the prefix length in the BGP update message is longer than the maximum length permitted in the RV record. By default, a prefix that is marked Invalid is not advertised to any peer. It gets withdrawn from the BGP routing table if it was already advertised, and is not flagged as the best path or considered as a candidate for multipath.

Not Found prefixes are the prefixes that are not found in the RV database. By default, a prefix marked as Not Found is installed in the BGP routing table and will only be flagged as the best path or considered as a candidate for multipath if there is no Valid alternative.

5.3 Protection Against BGP Hijacks

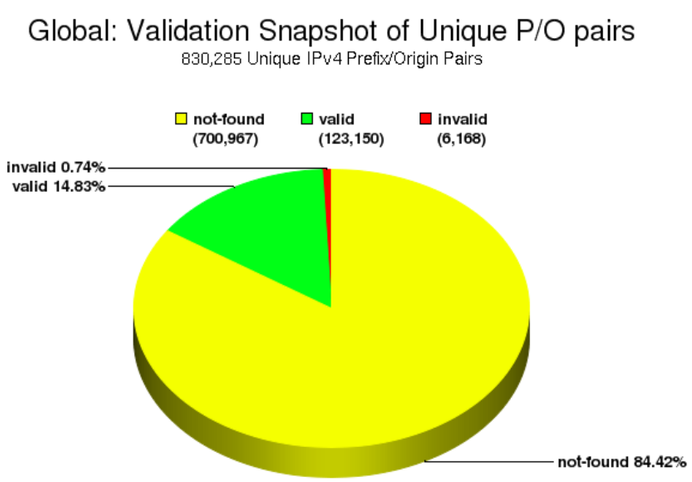
BGP origin validation protects against BGP leaks that are accidental announcements having no malicious purpose. A prefix is rejected if an origin AS in BGP UPDATE message does not match the AS in an RV record or the length of the prefix is longer than the allowed length. However, origin validation fails when an attacker announces a prefix with the spoofed AS number which is authorized to announce a prefix. The RPKI mechanism will not detect this; it simply trusts that the AS path is correct.

The BGPsec protocol defined in RFC8205 addresses this problem assuring that the entire path from the origin AS to the destination AS is valid. Each BGP speaking router on the path adds its local AS along with a prefix as well as the AS number of the receiving peer, it is going to send a BGPsec UPDATE message to. This information is signed by a private key and the router attaches a digital signature along with the hash of the public key to the BGPsec update message as BGPsec AS_PATH attribute. The BGPsec UPDATE message is sent to an eBGP neighbor. The router validates the signatures and if they are invalid, the route is rejected. If the signatures are valid, the router adds its own signature signed with its private key and containing local AS and the AS of the receiving neighbor, plus a hash of the public key. Every BGP speaker on the path validates all the signatures in a message to determine the authenticity of the path information contained in the BGPsec_Path.

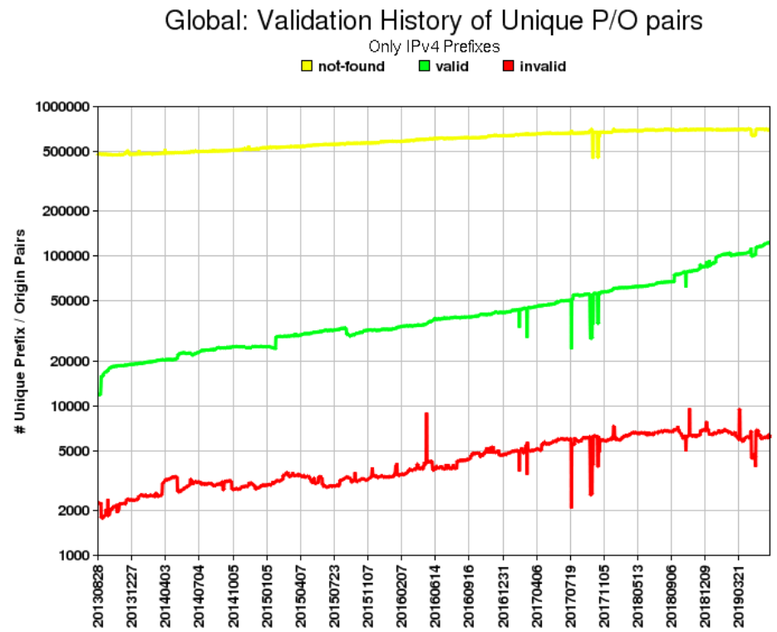
BGPsec is intended to be used to supplement BGP origin validation and when used in conjunction with origin validation, it is possible to prevent a wide variety of route hijacking attacks against BGP. The BGPsec router can talk with a non-BGPsec router but it must convert the BGPsec_Path to a regular AS_PATH, stripping

off all security information. However, to ensure that every AS on the path of ASes listed in the UPDATE message has explicitly authorized the advertisement of the route, there must be an unbroken path of BGPsec capable routers between an origin AS and destination AS. This is in contrast to prefix filtering and prefix origin validation where only the AS doing the filtering needs to deploy the security policy. The validation of digital signatures can be a CPU intensive task that may require an upgrade of existing hardware.

When running BGP it is important to not only protect your own network, but other networks as well. Unfortunately, we are still a long way from RPKI and BGP origin validation to be adopted on the Internet at large. The current state as well as the history of daily validation results of unique IPv4 prefix/origin pairs using global RPKI is depicted in Picture 14 and Picture 15.



Picture 14 - *The daily snapshot of validation results for unique Prefix/Origin pairs (https://rpki-monitor.antd.nist.gov/)*



Picture 15 - *The history of daily validation results for unique Prefix/Origin pairs. (https://rpki-monitor.antd.nist.gov/)*

The number of BGP leaks/hijacks, on the other hand, continues to grow. As per the BGPmon statistics, almost 2300 leaks/hijacks happened over the first 7 months of 2019.

Now, more than ever, proper BGP filtering remains of utmost importance.

NOCTION
NETWORK INTELLIGENCE

This ebook was brought to you by Noction