Orange International Carriers

BGP Best Practices for IP Transit Customer

This document outlines BGP configurations available to AS5511 IP Transit customers looking to add additional access circuits for IP Transit service performance and reliability optimization

16 March 2021
Version be84d190

Contents

1 Border Gateway Protocol best practices

The Internet today has grown into a worldwide network. People and companies rely on it for their day to day communication – from shopping and entertainment to business critical applications. As more services converge to IP transport, Internet reliability and quality performance becomes an extremely important objective for both Internet subscribers and Internet Service Providers (ISPs). While IP is used for Internet transport, Border Gateway Protocol (BGP) determines a path IP packets are transported on. BGP can steer traffic to a path of better performance and recalculate a path in case of a network failure.

This document outlines BGP configurations available to AS 5511 Transit customers looking to add additional access circuits for Internet service performance and reliability optimization. All mechanisms outlined below deal only with the traffic from Internet towards the customer network. Outbound traffic path to Internet should also be adjusted by tuning the routing decision of the customer routers. Customer network design and routing configurations are outside of the scope of this document.

2 BGP basics

- BGP is the path-vector protocol for routes exchange and best route calculation.
- The current version of BGP is 4, based on IETF standard RFC 4271.
- BGP speakers are manually provisioned and exchange routing information via a TCP connection. There is no auto-discovery in BGP.
- ISP and Internet subscribers' networks are usually identified by an Autonomous System (AS) - a unique identifier within the Internet. External BGP sessions are established between BGP speakers in different AS.
- BGP speakers exchange paths and their attributes. If BGP speaker receives multiple routes to the same destination it picks only one best route based the routes attributes. The easiest attribute to understand is AS_PATH. The path that traverses the least number of AS "wins." Other important attributes include NEXT_HOP, MULTI_EXIT_DISC (MED), LOCAL_PREF, ORIGIN, and COMMUNITY.
- BGP best path selection - BGP speaker starts with all available paths to a given destination and goes through a number of steps. At each step BGP speaker leaves only routes that satisfy certain criteria and stops when only one route remains. Note that vendors might add steps to the algorithm if not conflicting with the BGP RFC.
- An important aspect of BGP is a loop prevention mechanism. BGP speakers will not import routes that contain themselves in the AS_PATH.

BGP path selection algorithm steps	Description
1. Largest WEIGHT	Cisco proprietary step. Pick the routes with largest WEIGHT value.
2. Highest LOCAL_PREF	LOCAL_PREF does not get passed between different ASs.
3. Locally originated	Cisco proprietary step. Prefer the path that was locally originated via a network or aggregate BGP subcommand, or through redistribution from an IGP. Local paths sourced by network or redistribute commands are preferred over local aggregates.
4. Shortest AS_PATH	Count number of AS numbers in AS_PATH. An AS_SET counts as 1. The AS_CONFED_SEQUENCE and AS_CONFED_SET are not included in the AS_PATH length. Confederation hops ARE NOT counted as hops.

5. Lowest ORIGIN Type	Prefer the path with the lowest origin type: IGP < EGP < Incomplete - IGP ORIGIN for routes manually provisioned with bgp "network" command or when IGP is redistributed into BGP - EGP ORIGIN for routes learned from EGP protocol - Incomplete ORIGIN for routes learned via other means. This usually occurs when we redistribute a static route into BGP.
6. Lowest MED	MED is a hint to external neighbors about the preferred path into your AS. By default MED is only compared for routes from the same AS.
7. eBGP preferred over iBGP	eBGP routes are those learned via external BGP sessions (from different ASs). iBGP routes are learned via internal BGP sessions (same AS).
8. Lowest IGP metric	IGP metrics are the internal routing protocols that BGP uses to reach the next-hop ip address
9. Oldest route	Prefer the path that was received first.
10. From speaker with lowest Router ID	Prefer the route coming from the BGP speaker with the lowest router ID
11. From speaker with lowest IP	Prefer the path coming from the lowest neighbor IP address

Note that prefixes smaller then /24 are usually blocked by Internet ISPs. To allow routing for smaller prefixes, Customer Premises Equipment (CPE) has to advertise smaller prefixes with "no-export" COMMUNITY, and advertise aggregate prefix larger or equal /24 covering all small prefixes over each eBGP session.

3 Multiple access IP Transit circuits to AS 5511

For customers with a single access circuit a simple static routing would be sufficient and BGP routing is optional. Customer with more than one access circuit are provisioned with BGP routing for automatic fail-over. AS 5511 supports several configuration options: eBGP multipath, eBGP multihop, Link bonding, Active-Backup, and Active-Active.

3.1 eBGP multipath

When additional capacity is required between the same pair of Customer Premises Equipment (CPE) and a AS 5511 border router (BR) additional circuit(s) could be provisioned as an alternative to a single link upgrade. For each parallel circuit we provision a separate eBGP session. All eBGP sessions are one hope sessions using access interface IP.

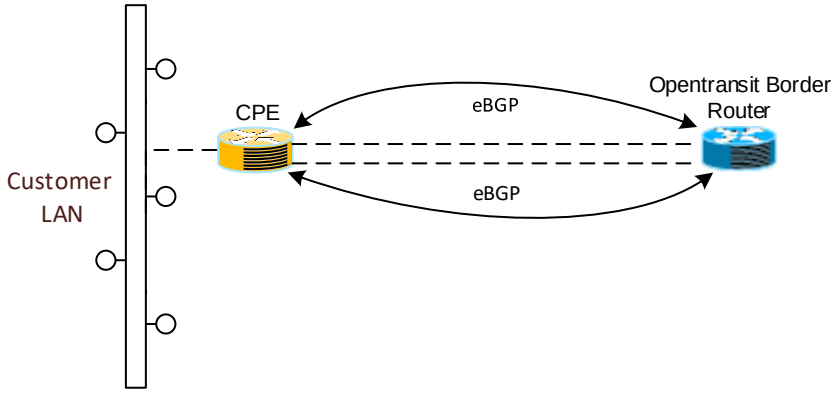

Figure 1. eBGP multipath

3.2 eBGP multi-hop

Another option to provision multiple access circuits between same CPE and a AS 5511 BR is “eBGP multi-hop”. In this case a single eBGP session is established between Loopback addresses of CPE and BR. In addition CPE is provisioned with one static route per link towards BR loopback pointing to the BR end of the link. BR is provisioned with one static route per link towards CPE loopback pointing to the CPE end of the link.

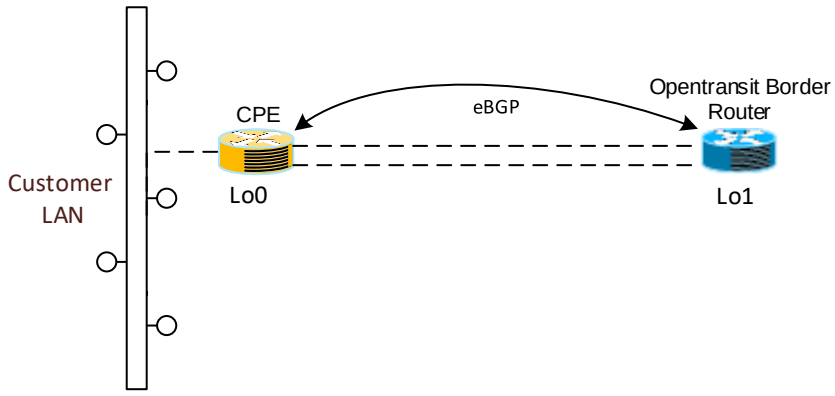

Figure 2. eBGP multihop

3.3 Links bonding

3rd option to provision multiple access circuits between same CPE and a AS5511 BR is "link bonding". Multiple Ethernet circuits could be bonded together using LACP. A single logical interface at each side represents and manages multiple circuits. In this case a single one-hop eBGP session is provisioned over this logical interface. Non-Etherent links can also be bonded with MLPPP using similar approach.

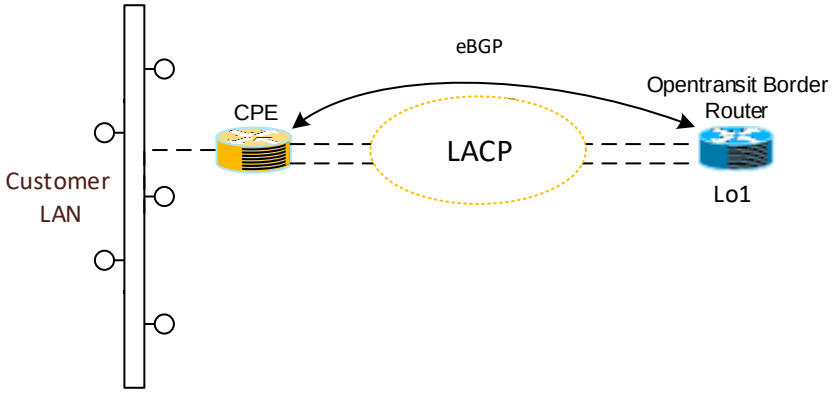

Figure 3. Link Bonding

3.4 Active-Backup

For additional router and site redundancy a customer can order additional circuit from the same or different CPE. Each circuit is provisioned with a separate eBGP session. Customer designates one circuit as Active by CPE advertising all prefixes with lower MED attribute. CPE advertises same prefixes over the second circuit with higher MED and the second circuit becomes a Backup.

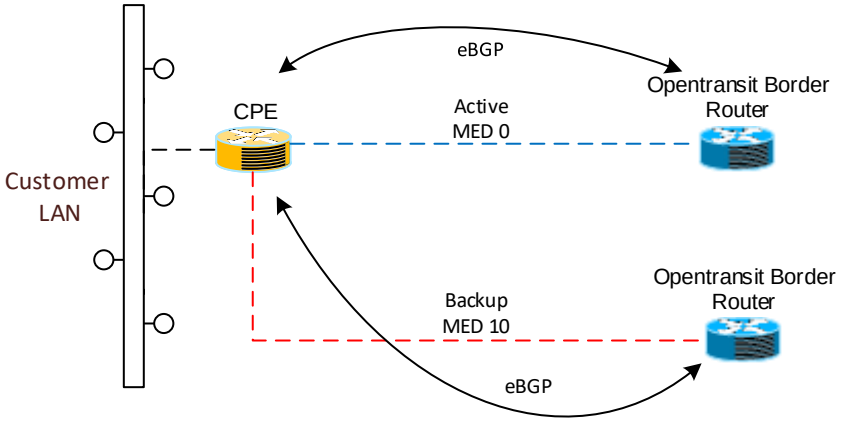

Figure 4. Active-Backup via MED

Alternative method to achieve the same result is for CPE(s) to advertise all prefixes over one session with COMMUNITY 5511:90. The AS 5511 BR will automatically translate COMMUNITY=5511:90 to LOCAL_PREF=90. In this case the first circuit becomes Active and the second – Backup.

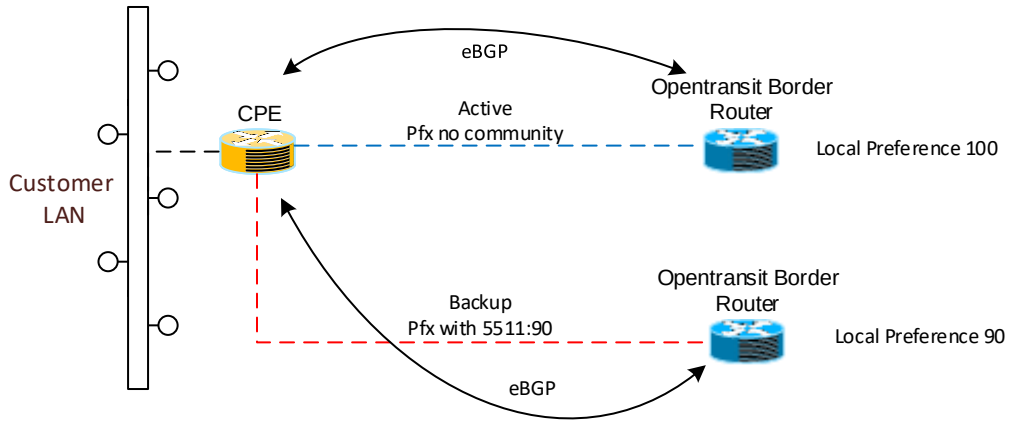

Figure 5. Active-Backup via LOCAL_PREF

In a more general case, a customer can designate access circuits as Active + Backup(s) with different priorities by advertising prefixes over corresponding eBGP sessions with COMMUNITY attribute 5511:70, 5511:80, 5511:90, and no community. These communities will be automatically translated by the AS 5511 BR to LOCAL_PREF 70, 80, 90, and 100 respectively. The higher the LOCAL_PREF the higher the circuit priority.

Please note that by default the AS 5511 marks all routes from customer with LOCAL_PREF = 100, from peers - with LOCAL_PREF = 85. As a consequence customer routes received from peers will be preferred over the same routes received from the customer with COMMUNITY 5511:70 or 5511:80.

3.5 Active-Active

With “Active-Active” setup each access circuit connect to the same or different CPEs and is preferred for a traffic to a sub-set of customer prefixes – in our example Prefix1 and Prefix2. This configuration can be beneficial if destinations covered by Prefix1 are close to CPE terminating circuit 1 and destinations covered by Prefix2 – close to CPE terminating circuit 2. Each access circuit still has to be scaled to carry 100% of the traffic if we want to avoid congestion and packet loss in case of one circuit failure. Another reason for such configuration might be a need to load-balance traffic over multiple access circuits because each circuit alone cannot handle 100% of the traffic – in this case we have to expect congestion and performance degradation in case of a circuit failure.

To achieve “Active-Active” setup, CPE advertises Prefix1 with MED=0, Prefix2 with MED=1 over the first circuit, and Prefix1 with MED=1, Prefix2 with MED=0 over second circuit. This makes first circuit preferred for traffic to Prefix1, and second circuit preferred for traffic to Prefix2.

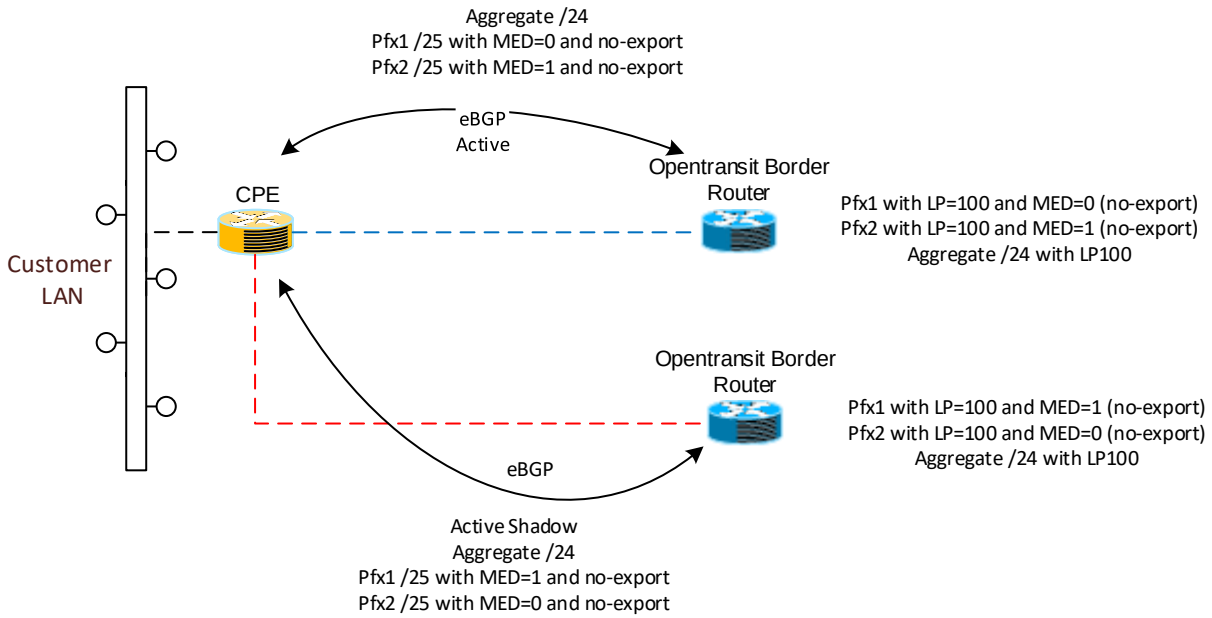

Figure 6. Active-Active via MED

Same functionality can be achieved via COMMUNITY to LOCAL_PREF translation mechanism explained in previous sections.

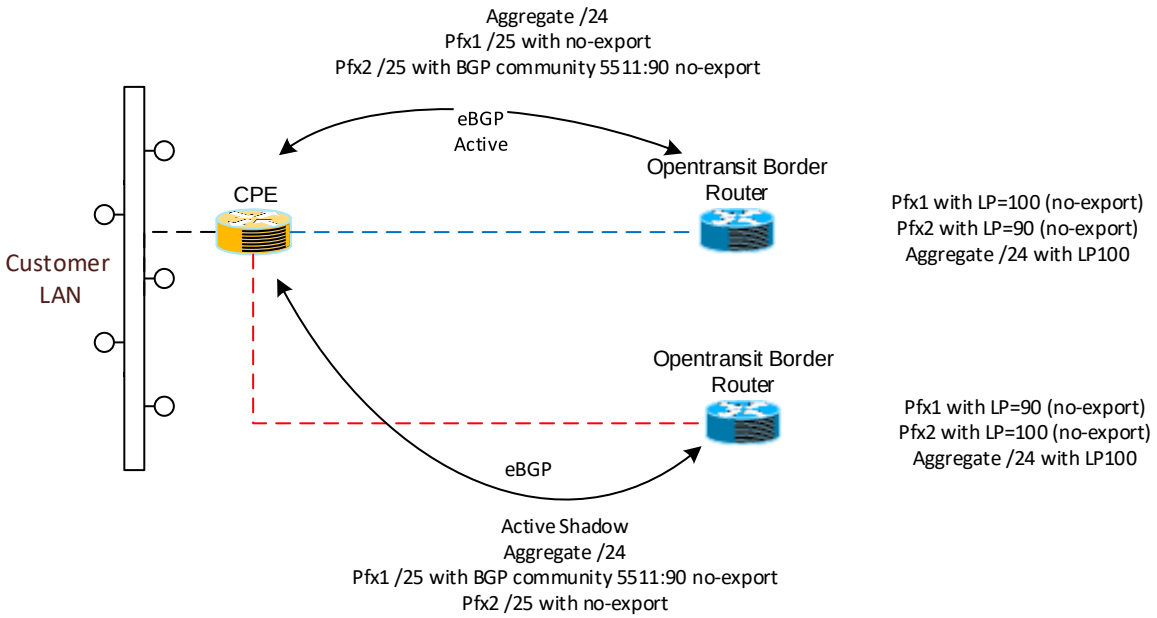

Figure 7. Active-Active via LOCAL_PREF

4 IP Transit access circuits to different ISPs (multi-homing)

A customer can connect to two or more ISPs. This setup, commonly known as BGP multi-homing, provides additional protection against ISP network outages, but adds complexity in managing traffic flow via the access circuits.

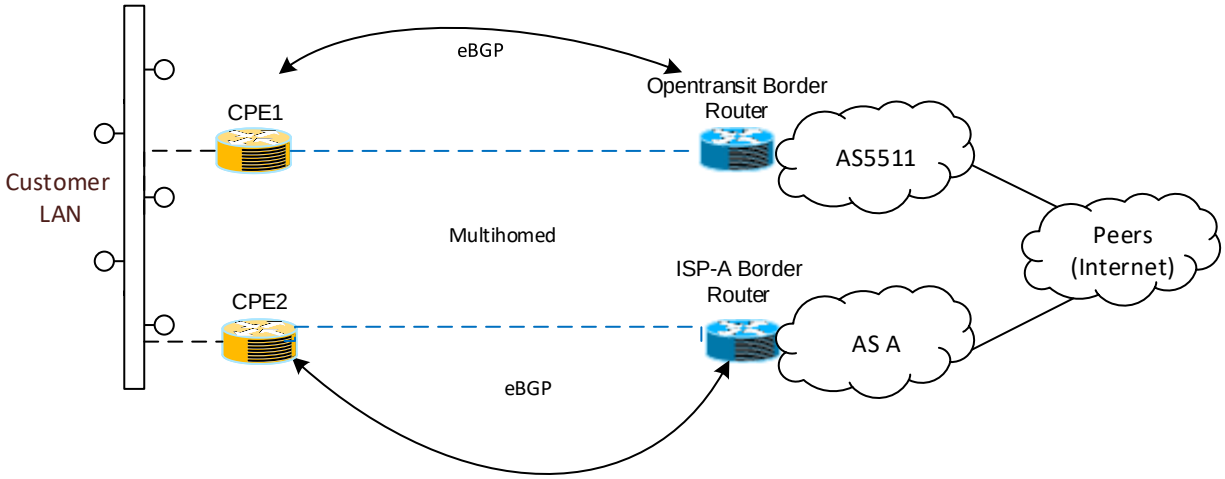

Figure 8. Multi-homing

In multi-homing scenario a customer typically faces a decision to accept either partial or full Internet routing table from an ISP. The AS 5511 typically advertises full routing table or a default route. Partial routing table can be achieved on CPE by filtering out a portion of the received prefixes based on IP addresses and/or communities attached by ISP. Default route is not allowed to propagate outside of the customer's routing boundaries.

If a customer advertises the same prefix to two or more ISPs, BGP path selection algorithm by default favors the ISP with the larger customer base and peering connectivity, which translates into a fewer number of networks to a given destination or shorter AS_PATH. The closer the destination, the better usually is the quality and reliability of the network path. But for various technical or commercial reasons a customer can fine-tune BGP exchange to override the default behavior. When connecting to a single ISP, as we saw in previous sections, a customer can manage access circuits traffic flows by modifying MED or using COMMUNITY to LOCAL_PREF translation mechanism. But MED and LOCAL_PREF attributes are non-transitive, won't be propagated beyond the ISP's AS, and will have no effect on a routing decision beyond this ISP. To overcome this limitation in multi-homing scenario, CPE can "pad" the AS_PATH attribute for a given prefix – CPE adds its own AS number to the AS_PATH two or more times instead of the default one and makes a corresponding circuit less preferable for this prefix. AS_PATH is a mandatory and transitive BGP attribute therefore it would be propagated across AS boundaries.

Additionally AS 5511 IP Transit customers can attach special COMMUNITIES to the advertised prefixes and signal certain action on the AS 5511 side. The list of such COMMUNITIES is provided below.

5 AS 5511 BGP communities

IP Transit customers running BGP can advertise COMMUNITIES with their routing updates to trigger a certain routing adjustment on the AS 5511 network, such as changing priority for a route via a certain peer, or deflecting a DDoS attack (requires additional setup), or "black-hole" a traffic to a given host prefix.

.

COMMUNITY	Reaction on OTI router	Result
5511:0	Set NEXT_HOP=Null0	Traffic to host prefix is dropped
5511:70	Set LOCAL_PREF=70	Default LOCAL_PREF for customer routes is 100, for peer routes - 85.
5511:80	Set LOCAL_PREF=80	Default LOCAL_PREF for customer routes is 100, for peer routes - 85.
5511:90	Set LOCAL_PREF=90	Default LOCAL_PREF for customer routes is 100, for peer routes - 85.
65535:65281 or "NO_EXPORT"	Not advertised outside of OTI	As defined in RFC1997. Do not advertise outside of OTI.
65535:65282 or "NO_ADVERTISE"	Not advertised to any BGP speaker	As defined in RFC1997. Do not advertise to any BGP speaker
65535:65284 or "NOPEER"	Not advertised to OTI peers, advertised to OTI customers only	As defined in RFC3765. Do not advertise to OTI peers, but advertise to direct customers.
5511:100x	Special treatment when announced to all American OTI Peers	x=0 - do not announce to these peers; x=1-announce to these peers with 1 prepend; x=2–announce to these peers with 2 prepends.
5511:200x	Special treatment when announced to all European OTI Peers	x=0 - do not announce to these peers; x=1-announce to these peers with 1 prepend; x=2–announce to these peers with 2 prepends.

5511:300x	Special treatment when announced to all Asian OTI Peers	x=0 - do not announce to these peers; x=1-announce to these peers with 1 prepend; x=2–announce to these peers with 2 prepends.

Special treatment when announced to a specific OTI Peer with ASN

5511:150x	AS209 Centurylink	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:151x	AS2914 NTT/Verio	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:152x	AS7018 ATT/Worldnet	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:153x	AS6461 Abovenet/MFN/Zayo	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:154x	AS6453 VSNL/Tata/Teleglobe	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.

5511:155x	AS1239 Sprint	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:156x	AS1299 Telia	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:157x	AS3561 Savvis (exC&W)	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:158x	AS3549 Global Crossing	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:159x	AS3356 Level3	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:160x	AS701 UUNET-US/Verizon	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:161x	AS3320 DTAG	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.

5511:162x	AS286 KPN	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:163x	AS174 Cogent	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:164x	AS3257 Tiscali	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.
5511:165x	AS4134 China Telecom	x=0 - do not announce to ASN; x=1 - announced to ASN with 1 prepend; x=2 – announced to ASN with 2 prepends.

6 RPKI filtering

The AS 5511 network is in the process of filtering inbound **RPKI invalids**, whether such an advertisement is received from an IP Transit customer, or a peer.

If a prefix advertised by an IP Transit customer is covered by a ROA, **consistency between the BGP advertisement and the covering ROA(s) should exist and be maintained over time**.

Failure to do so will cause prefixes assigned a "RPKI invalid" state by our border routers to be discarded.