The Comprehensive Guide to BGP

HANDBOOK

catchpoint™

TABLE OF CONTENTS

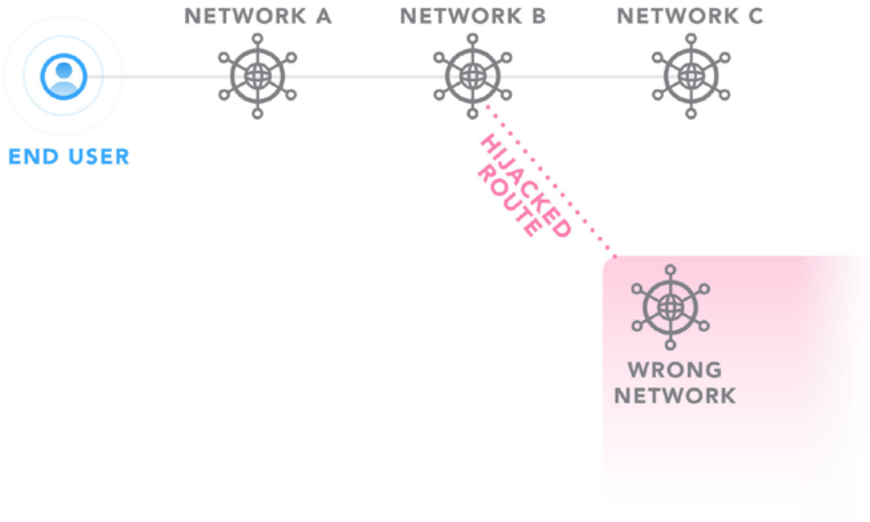
The Border Gateway Protocol (BGP) turned 30 years old this year, making it one of the most long-lasting, widely-used protocols ever deployed in the Internet. BGP was initially conceived in January, 1989 by Yakov Rekhter (IBM) and Kirk Lougheed (Cisco) on two napkins during the 12th IETF conference in Austin, Texas.

Curiously enough, BGP was conceived as an interim solution to overcome the infeasibility of using the existing Exterior Gateway Protocol (EGP) with the increase in complexity for connectivity between Administrative Domains.

Thirty years passed, and the interim solution became one of the pillars of the Internet architecture. Version 4 (the current version) was released in 1994, and since then it gets updated sporadically with new features and capabilities.

SOME BASICS

Before we dive into the history of BGP, let's go over some basics of what it is. The primary function of BGP is to manage how packets are routed across the internet through the exchange of routing and reachability information between edge routers. BGP directs traffic between autonomous systems (AS), which are network routers managed by a single enterprise or service provider. When an AS gets set up, it peers with other AS's to share IP prefixes, which are then shared with other AS's, and so on. In this way, when new prefixes are announced, they get propagated around the internet.



The biggest problem, however, is that BGP is extremely vulnerable to both malicious attacks and human error. There are roughly 65,000 AS's that make up the global internet, and little to no oversight for how each AS peering filters must be configured. This means that if a new, bogus route (aka a bogon prefix) is announced (either through intentional hijacking or just a typo) it sends traffic to the wrong network, and can spread like wildfire across the internet.

A LITTLE BIT OF HISTORY

Some background is required to better understand the crucial role that BGP played in the history of the Internet. In 1989, the Internet as we perceive it today was just moving its first steps. The commercial use of the Internet was still forbidden (the restriction was lifted in 1995 with the decommission of NSFNET), but commercial ISPs were sprouting and offering network access to end users, and the commercial use of the Internet was no longer a taboo topic.

When BGP was firstly standardized in June, 1989, the long-running ARPANET was just being decommissioned (February 28, 1989), TCP/IP was being used to interconnect different networks from remote countries, and the Internet was about to move from its centric architecture to a more distributed architecture, without a clearly defined backbone. Curiously, the requiem to ARPANET by Vinton G. Cerf was performed in the very same IETF meeting where BGP was just being announced to the world.

Up until then, the so-called Internet gateways were exchanging net-reachability information via the Exterior Gateway Protocol (EGP). EGP was conceived for an Internet composed by a core AS and multiple other smaller AS's directly connected to that core, and it totally relied on having a tree-structured topology of AS's, without cycles.

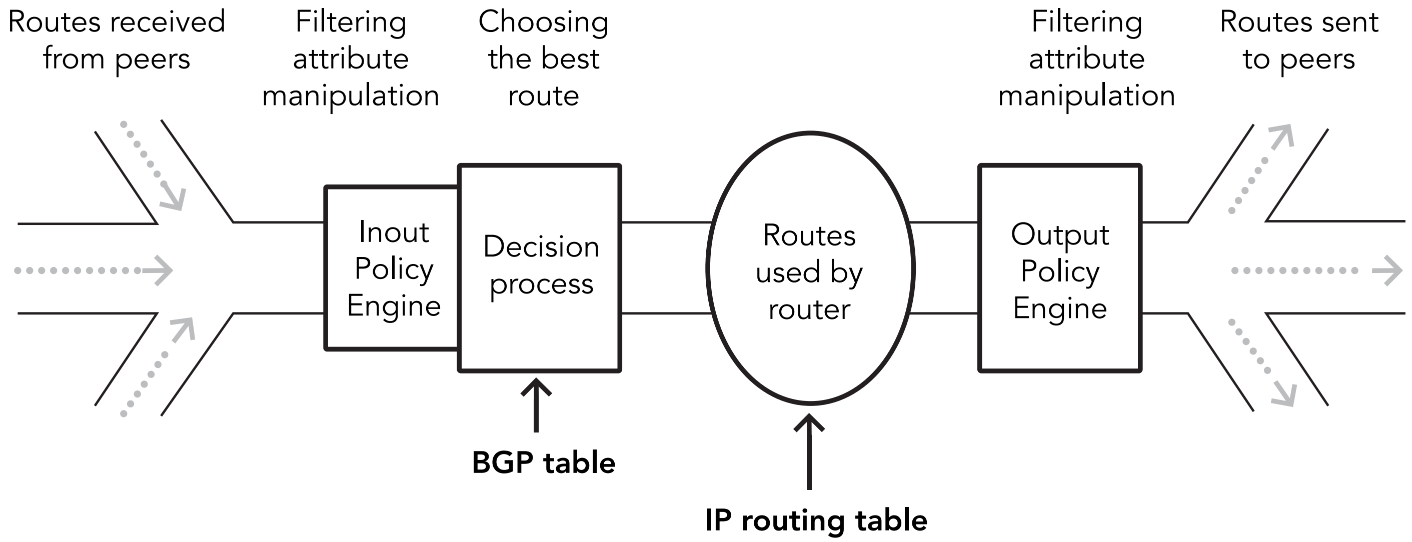
Although these limitations were bearable in an early stage Internet where stub gateways were talking to each other via its ARPANET backbone, with the advent of commercial entities and multiple backbones (such as NSFNET), its inadequacies became more and more pronounced – not to mention the impossibility to create policy-based routing, which is the key of success of BGP.

BGP: THE TWO-NAPKIN PROTOCOL

BGP is still a path vector protocol like its predecessor EGP, but it was conceived foreseeing a peer-to-peer environment where AS's could exchange routing information without relying either on a priori topology knowledge or on a core AS. With the introduction of BGP, the concept of AS has also been changed and re-defined. In the last BGP version, an AS "is considered to be a set of routers under a single technical administration, using an interior gateway protocol (IGP) and common metrics to determine how to route packets within the AS, and using an inter-AS routing protocol to determine how to route packets to other AS's."

The basic piece of routing information that AS's exchange with each other is called route. A route is composed by a set of destination IP networks paired with set path attributes, which describe the path toward the destinations. "This information is sufficient to construct a graph of AS connectivity from which routing loops may be pruned and policy decisions at an AS level may be enforced."

To guarantee the reliability of transmission, BGP is encapsulated into a TCP connection, meaning that two routers that want to establish a BGP session must have prior IP reachability. After establishing a TCP connection, the two routers – hereafter called BGP peers – agree on the parameters to use in the BGP session via BGP open messages, and then start exchanging routes. These routes can be generated by the peer itself or they can be learned by the peer via other BGP sessions, and each of them is announced via BGP update messages.



The figure above summarizes the BGP process each AS applies when receiving a route from another peer. Whenever a route is received from an AS, the route is subject to a filtering process where it can be discarded or accepted and, if required, its path attributes are manipulated. Then a BGP decision process is applied to select the best route for each IP destination network, since an AS may receive multiple routes toward the

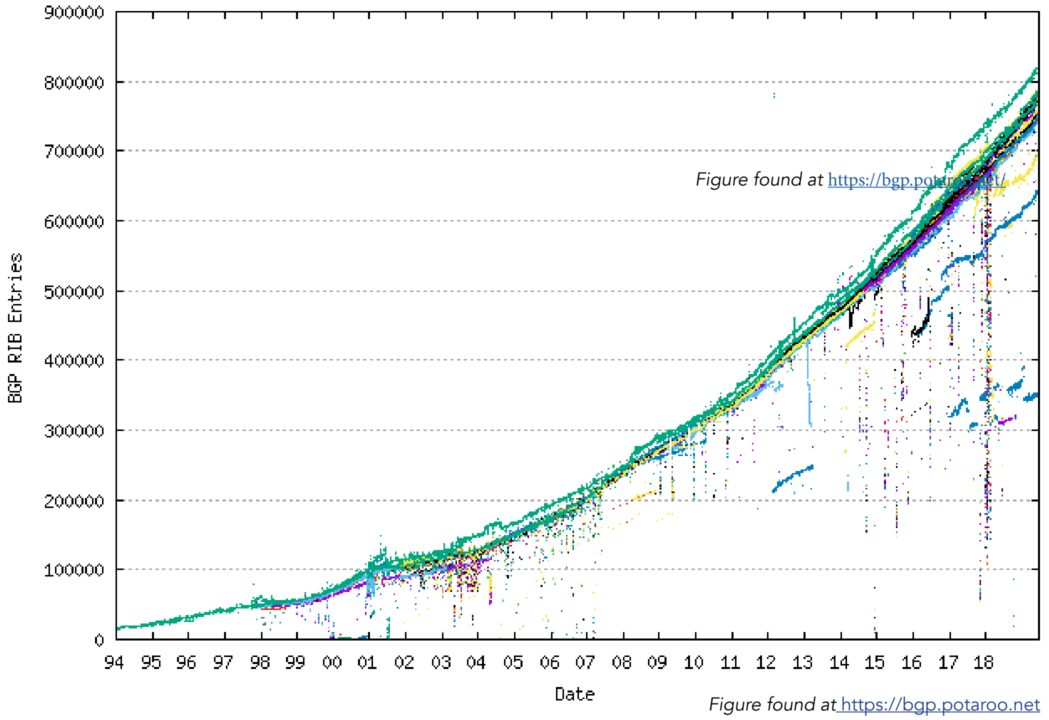


Figure found at https://bgp.potaroo.net

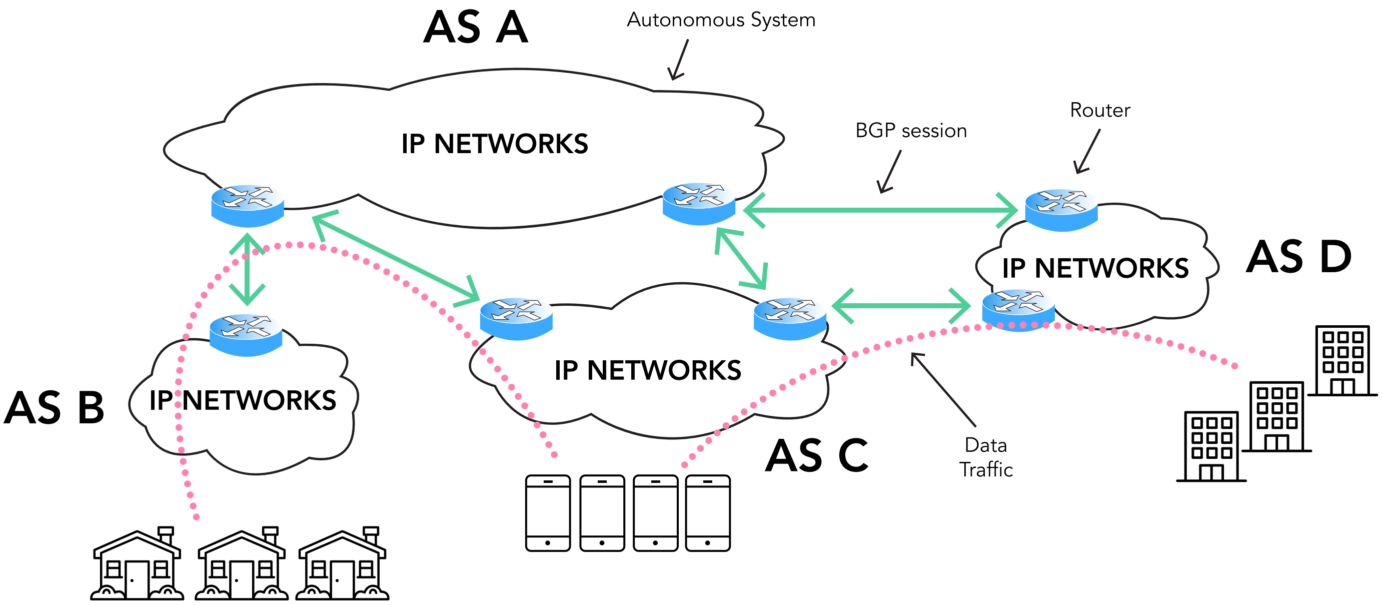
same IP network from different peers.
The BGP decision process is composed of a sequence of steps that allow the AS to choose the best route by analyzing the path attributes of each of the candidates, in order to apply criteria that range from pure commercial (e.g. prefer a cheaper provider over the other) to technical reasons (e.g. transit traffic to reach a destination via the smallest number of ASes). Each best route is then installed in the routing table of the router and used to forward traffic. Eventually, after a proper attribute manipulation, each best route is propagated to the all other BGP peers, or a subset of them depending upon the output filtering process applied.

Since the early days of deployment of BGP, the Internet grew widely in size and shape. Nowadays, the Internet is composed of about 65,000 AS's that exchange routing information related to about 800,000 IPv4 networks and about 70,000 IPv6 networks. However, due to the distributed architecture of the Internet, it is impossible to determine the number of BGP sessions established among AS's in the wild.

X-RAYING BGP

Two AS's exchange routing information by establishing BGP session(s) between pairs of routers running a BGP daemon, namely BGP speakers. After establishing a BGP session, the BGP speakers start exchanging the set of network prefixes that they either received from other AS's or that they already possessed. Each network destination exchanged is paired with a set of attributes that describes the characteristics of the path to reach that destination, forming what is called a route. Eventually, a BGP speaker will receive the routes related to all the Internet destinations. With these routes, the BGP speaker can forward traffic towards any intended destination.

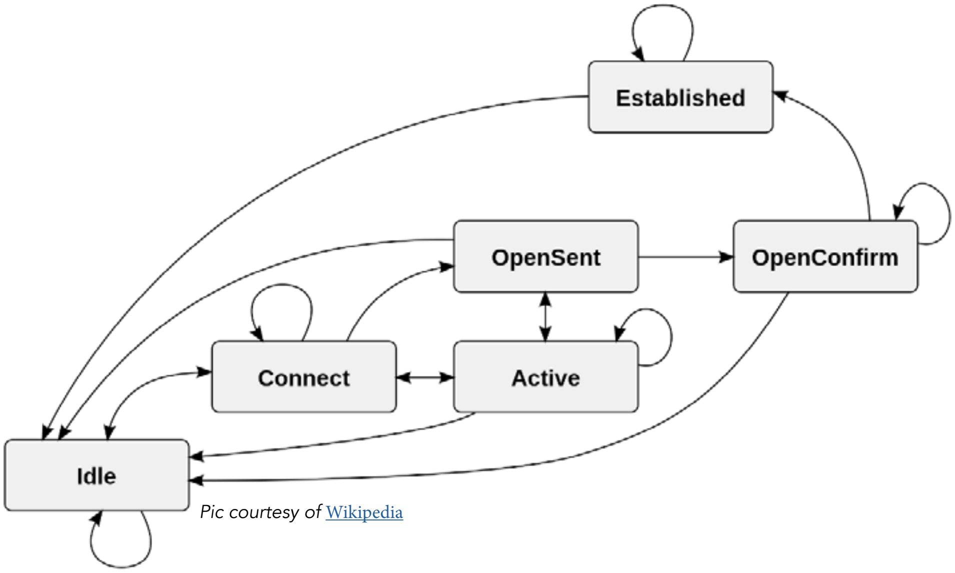
The set of attributes associated with each route enables a BGP speaker to implement routing policies which may reflect either commercial agreements it has with its neighbors or technical considerations. This flexibility is one of the key factors that allowed BGP to become the standard de-facto routing protocol of the Internet.



BGP OPERATIONS

BGP protocol is quite unique in the family of the routing protocols. Its most relevant peculiarity is that it relies on TCP (port 179) to guarantee the ordered and reliable exchange of protocol messages. This is because – unlike other routing protocols – there is no peer discovery process, and each peer is statically configured by the network administrator. Indeed, BGP is conceived to be an inter-AS protocol where peers should have quite a large degree of stability, thus making the discovery process useless. Therefore, it is a sine qua non condition for two BGP speakers to have IP reachability.

The process of establishing a BGP session between peers is performed via a simple finite state machine (FSM), which is described in the original RFC and can be summarized in the below figure:



Pic courtesy of Wikipedia

The process can be simplified as follows:

- Each BGP speaker starts in *Idle* state. This is a transient state where the BGP speaker initializes the required resources for the connection, and where it starts to listen for TCP attempt connections on port 179 and, at the same time, attempts to connect to the other BGP speaker via TCP on port 179.

- Once these steps are performed, the BGP speaker moves to the *Connect* state, where it sets a timer and waits for the TCP connection to be completed. If the *ConnectRetryTimer* expires, the TCP connection is dropped, and the timer resets while still listening for incoming TCP connection attempts. If the TCP connection fails, the BGP speaker moves to the *Active* state.

- The *Active* state is another transient state where the BGP speaker basically stops to actively attempt to connect to the other party and just listen for incoming TCP connection attempts. Once the *ConnectRetryTimer* expires, the BGP speaker goes back to *Connect* state.

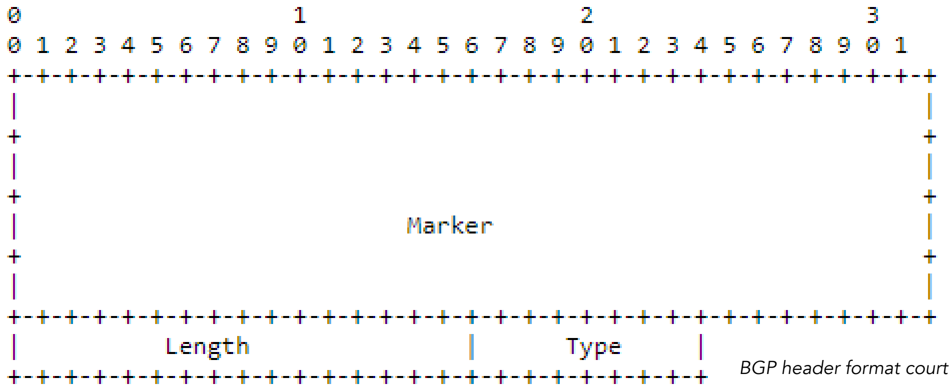
- If the TCP connection is successful either in *Connect* or Active states, the BGP speaker must send an *OPEN message* containing a list of its capabilities, moving to the *OpenSent* state.

- Once in *OpenSent* state, a BGP speaker waits for an *OPEN message* from the other party, and if no error occurs, sends a *KEEPALIVE* message, moving then to *OpenConfirm* state. Otherwise, it sends a *NOTIFICATION message* and goes back to *Idle* state.
-
- Finally, in the *OpenConfirm* state the BGP speaker waits for a *KEEPALIVE* message or a *NOTIFICATION message* from the other party. If it receives a *KEEPALIVE message*, then it moves to the *Established state*, otherwise it moves back to *Idle*. In general, any error in any state cause the BGP speaker to move to Idle state.

- Once in *Established* state, each BGP speaker announces *routes* towards via *UPDATE* messages to allow the other party to reach those destinations. The amount of destinations announced strongly depend on the agreement between network operators.

MESSAGES

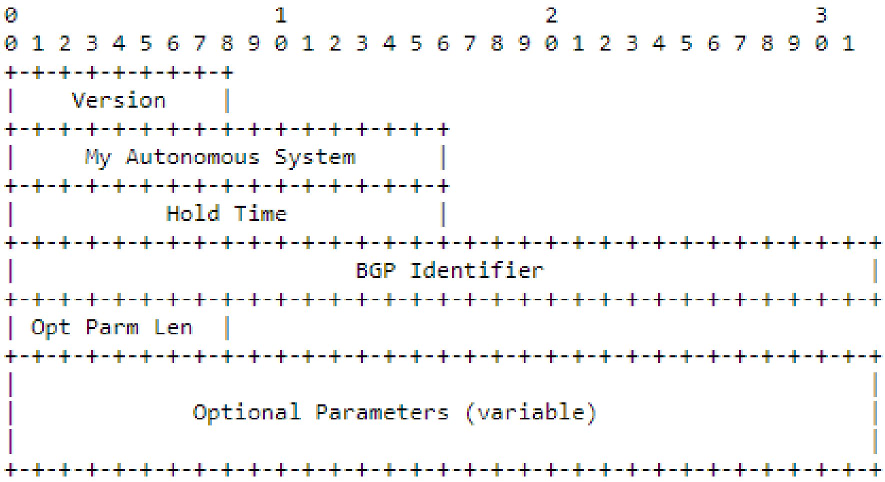
The evolution of the FSM described above is regulated via the exchange of BGP messages. Each BGP message starts with a common BGP header composed by 19 bytes and encoded as follows:



```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                                                               |
+                                                               +
|                           Marker                              |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Length               |      Type     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

BGP header format courtesy of IETF

The *Marker* is a 16-byte field set all to one. This field is included for compatibility with older BGP versions and has no specific semantic in the current BGP version. The *Length* field contains the length of the BGP message, header included. The original RFC specifies that the maximum length of a BGP message is 4096 bytes despite the field size. This value was considered to be more than enough for the protocol requirements. Finally, the *Type* field contains the type of the message. The most common BGP message types are four: *OPEN (1), NOTIFICATION (2), KEEPALIVE (3),* and *UPDATE (4).*



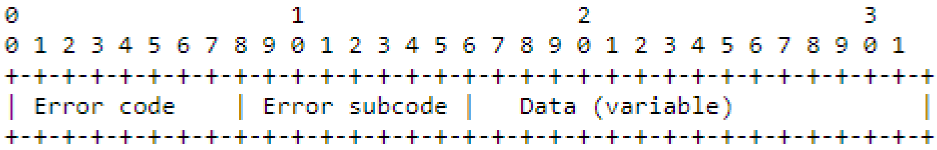
```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+
|    Version    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     My Autonomous System      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Hold Time           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         BGP Identifier                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Opt Parm Len  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|             Optional Parameters (variable)                    |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Open message format courtesy of IETF

1. The most important type of messages in the initial setup phase of a BGP session are the ***Open messages.*** As detailed in the FSM, these messages are used by the two BGP speakers to inform each other about the parameters they propose to use for the BGP session and inform the other party about their *capabilities*. Capabilities are Optional Parameters describing which BGP extension each speaker supports, like the support for four-octet AS numbers, the support of multiple protocols in BGP (e.g. IPv6), and the support for multiple paths. If two BGP speakers share a common capability, they will be automatically enabled to exploit the capability new features in the BGP session, like announcing each other's IPv6 routes.

In addition to capabilities, Open messages carry the current *Version* of BGP (set to 4 since 1994) and some information about each of the peer, like the self-explanatory *My Autonomous System* field and the *BGP identifier* field, where it is encoded one of the IPv4 addresses belonging to the announcing BGP speaker.

Another important mandatory parameter found in these messages is the *Hold Time*, which regulates how long the BGP session can stay up without the exchange of any protocol message. This parameter is crucial to avoid the reset of the session upon a temporary network failure; however, its value cannot be too high otherwise it could take too long for a speaker to realize that the session is no longer available. The BGP specification suggests using 90 seconds for that value.

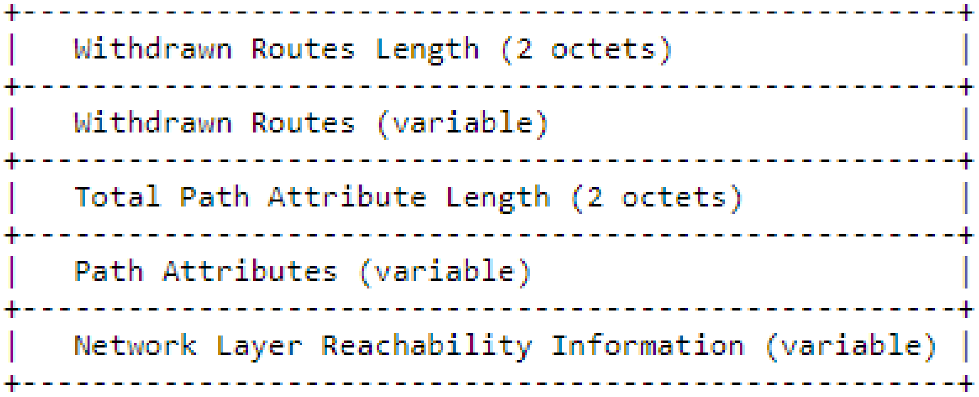


```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Error code    | Error subcode |   Data (variable)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

NOTIFICATION message courtesy of IETF

2. ***Notification messages*** are quite the opposite of the *Open messages*. They are triggered whenever one of the BGP speaker incurs in an error (for any reason). In these cases, the BGP speaker notifies the other party about the type of error it has experienced, just before tearing down the BGP session. The possibility was recently introduced to add some text to this message to show some human-readable information in the log of the other BGP speaker.

3. ***Keepalive messages*** are empty BGP messages, composed by the header and without any payload. These simple messages are used to acknowledge decisions (like in the setup phase) and to keep the session up in absence of routing information exchanged by the two BGP speakers.



```
+-----------------------------------------------------+
|   Withdrawn Routes Length (2 octets)                |
+-----------------------------------------------------+
|   Withdrawn Routes (variable)                       |
+-----------------------------------------------------+
|   Total Path Attribute Length (2 octets)            |
+-----------------------------------------------------+
|   Path Attributes (variable)                        |
+-----------------------------------------------------+
|   Network Layer Reachability Information (variable) |
+-----------------------------------------------------+
```

UPDATE message format courtesy of IETF

4. Finally, there are the *Update messages*. Those are the messages that carry the routing information that AS's exchange each other. An *Update* message can be thought as composed by three main parts: *Withdrawn Routes, Path attributes* and *Network Layer Reachability Information (NLRI).*

Withdrawn routes and *NLRI* fields are quite straight-forward. They contain the subnets that are the subject of the route carried by the UPDATE message. If the subnet is among the *Withdrawn routes*, it means that the BGP speaker has no more routes involving that specific subnet. If the subnet is among the *NLRI*, then it means that the BGP speaker found a new route to for that specific subnet, whose characteristics are described in the *Path attributes* field. This can either mean that a new subnet has been announced in the Internet, or that an already existing subnet could be reached via a different route.

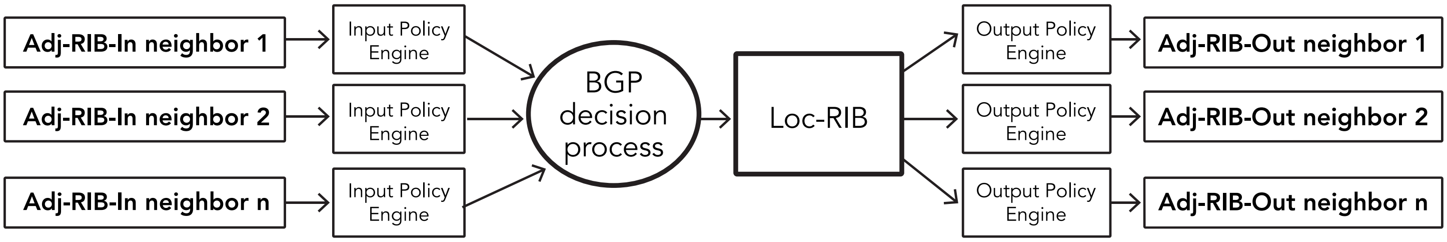
The *Path Attributes* field contains a set of attributes which describe the path toward the destinations contained into the *NLRI* field. There are many different attributes, each with its own role and format. The original RFC mandates that the *Path Attributes* field must be present if the NLRI field contains at least one destination, but only a few *Path Attributes* are required to be present:

- *AS_PATH*: the list of AS's that must be traversed to reach the destination via the given route
- *ORIGIN*: the origin of path information
- *NEXT_HOP*: the IP address of the router that should be used as next hop towards the given destination

ROUTE ELABORATION

Once the session is established, the two BGP speakers announce *Update* messages to each other to advertise their reachability to the other party. Since BGP speakers could be connected to multiple BGP speakers – possibly belonging to different AS's – and thus could receive multiple routes toward the very same destination, each BGP speaker must run a decision process to select the best route for each subnet received.

This is called *BGP decision process*, and that’s exactly where BGP conveys its flexibility.
More in detail, when a BGP speaker receives a BGP update message from a neighbor, it stores each route advertised into a table dedicated to the neighbor, called *Adj-RIB-In*.



Route elaboration process

Once a route has been installed into the Adj-RIB-In, it is checked against a filter to decide whether it can be accepted or not. The ingress filter is completely customizable by the network administrator, which can decide to discard a route for a plethora of reasons. For example, a route could be discarded if the destination network was not expected to be received from that specific neighbor, or if it contains a specific path attribute value.

If a route is accepted, then it participates to the *BGP decision process* together with all the accepted routes toward the same destination learned from other neighbors. This process is conceptually composed of three phases, with some slight differences from router vendor to router vendor.

The first phase is triggered whenever an *Update* message has been accepted and consists in calculating a degree of preference for each route advertised in the message.

Once this phase is completed, the BGP speaker chooses the *best route* among all the routes available for each distinct destination in the message and installs each *best route* in the *Local Routing Information Base (Loc-RIB)*. The *Loc-RIB* is the table containing all the routes that the BGP speaker is using to route the traffic received from its neighbors.

The third phase is triggered once the *Loc-RIB* has been modified. In this phase, each route that contributed to change the Loc-RIB is checked against neighbor-specific output filters. From there, it’s installed into the neighbor Adj-RIB-out and becomes ready to be advertised. Like the ingress filters, the output filters are completely customizable by the network administrator which could decide, for example, to exclude a neighbor to receive a route towards a specific destination.

BGP AND YOUR BRAND'S BOTTOM LINE

The Internet still works on the very same foundations that it started with back in the early 90s. A limited number of companies (e.g. CenturyLink, AT&T, Verizon) are offering transit via their worldwide backbone to a much larger number of companies. Among them, we have eyeball networks such as most of national telcos (e.g. Telecom Italia, British Telecom), regional providers (e.g. Apuacom), and CDNs (e.g. Comcast, Sky), all aiming at providing the best performances to end users – even if from different perspectives – via their interconnections, often established on Internet eXchange Points (IXPs).

Differently from the last few decades, new big players (e.g. Google, Facebook, Netflix) recently joined this interconnection game with enough resources to mine their own foundation. These players decided to play on their own by creating their own worldwide infrastructure, interconnecting directly to as many AS's as possible and offering direct access to their content, thus totally bypassing third-party backbones.

Up to date, the scenario is composed by about 65k AS's, mostly regional, interconnected to each other via BGP and each enforcing its role via a specific set of import/export filter policies. Among these 65k AS's, only about twenty of them can reach the whole Internet destinations without purchasing transit from any other AS, forming the so-called Tier-1 club.

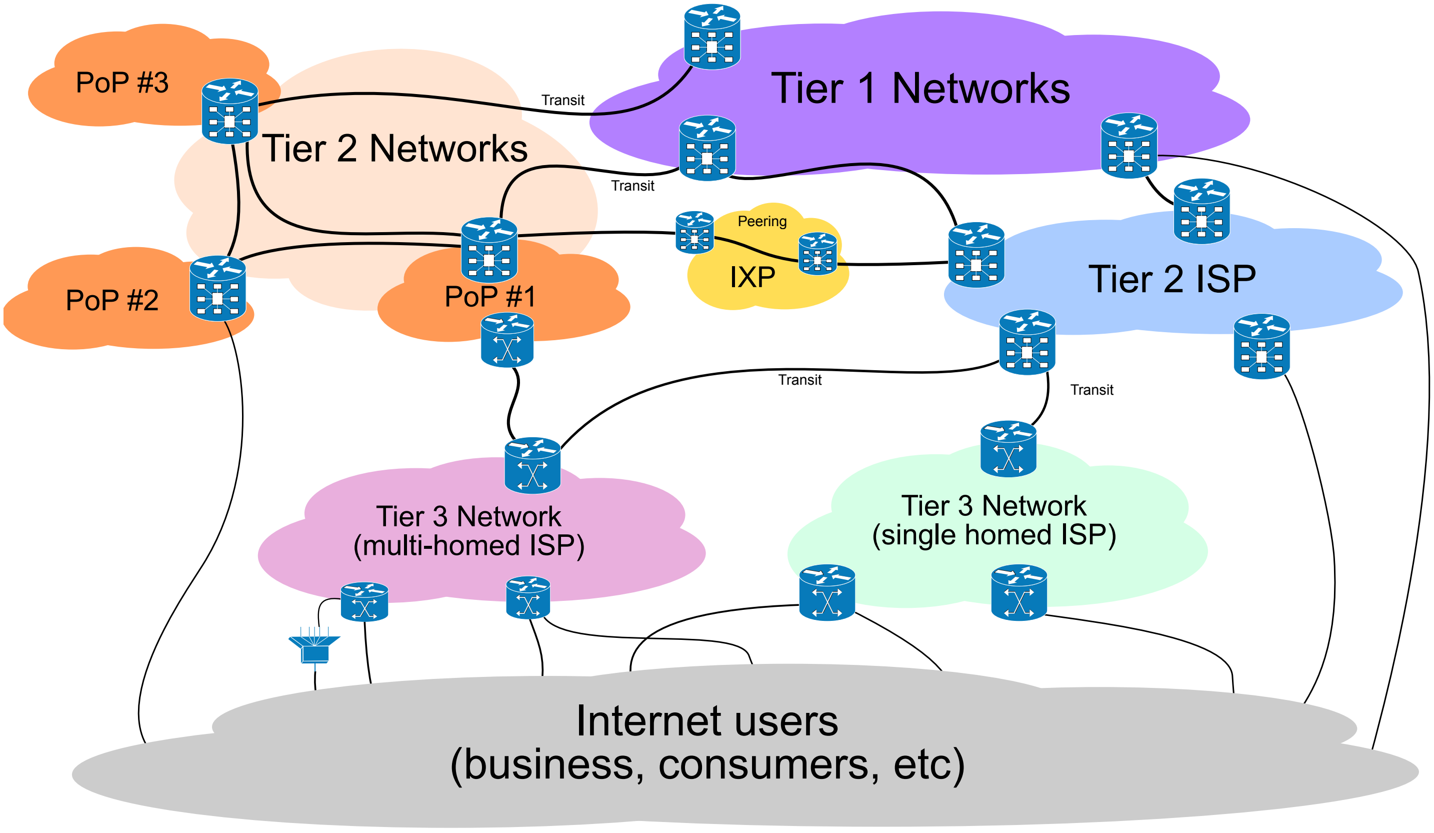
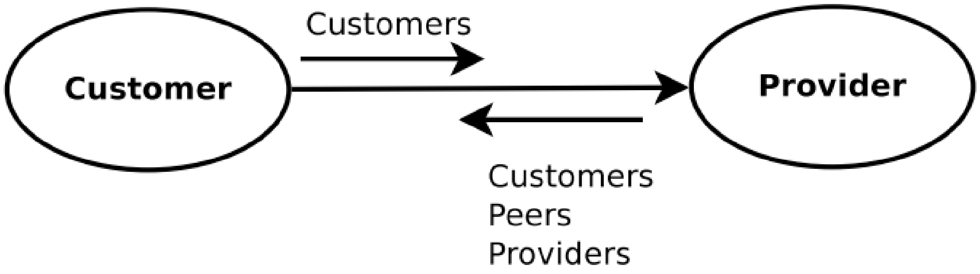


Image courtesy of Wikipedia

ECONOMIC RELATIONSHIPS AND BGP

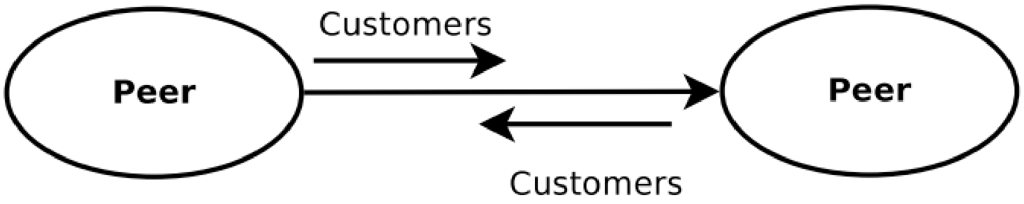
The Internet is a complex system made of interconnected AS's, each with its own role, market, and resources. Nevertheless, it is possible to roughly categorize the type of relationships existing between AS's in two broad categories, each identified by a BGP import/export filter policy.



The first relationship is **provider-to-customer (p2c)**, or **customer-to-provider (c2p)**, where one of the two AS's is providing transit to the whole Internet for the other AS. This type of relationship is often on a contract basis involving a fee paid by the customer to the provider, and it is established via private facilities.

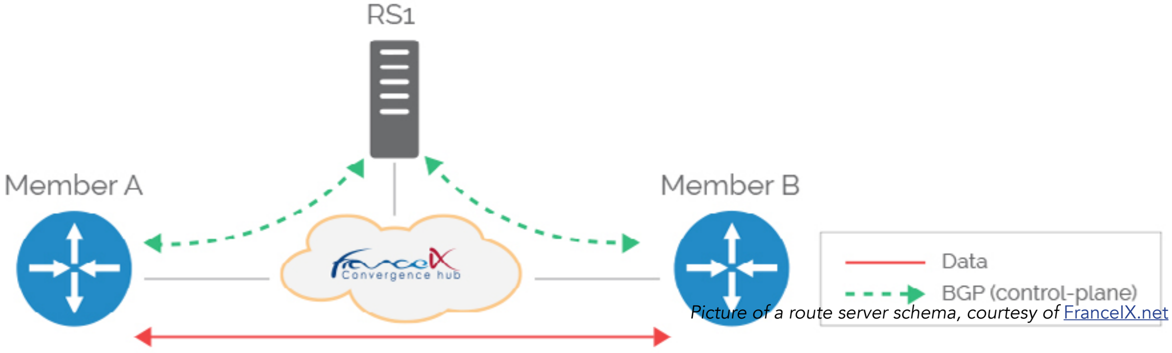
To fulfill its role, the provider announces to the customer every route required to reach the whole Internet. Depending on the agreement, this could consist of a single default route (0.0.0.0/0 in IPv4 and ::/0 in IPv6) or of a full routing table, which as of today consists of about 750k subnets in IPv4 and about 80k subnets in IPv6. On the other side, the customer will announce to the provider only its own routes and the routes received from its customers to allow the provider to use their interconnection to reach those destinations.

The second relationship is **peer-to-peer (p2p)**, where the two AS's decide to announce to each other the networks which each AS can reach without using any transit connection or any other p2p relationship. One of the main reasons behind these relationships is to keep traffic local as much as possible via public or private facilities, thus avoiding extra delays introduced by the transit connection which potentially can route the traffic via another country.



This type of relationship is typically settlement-free and established on IXPs and known as public peering. However, this is not always the case. Depending on the agreement made by the two AS's, p2p relationships can also involve a fee (paid peering) and/or can be established in private facilities (private peering). According to PCH survey, only very few p2p relationships are paid or private peering, and most of them are not formalized in any written document. Moreover, during the last year most AS's exploit route-server on IXPs to establish p2p relationships among them (multi-lateral peering).

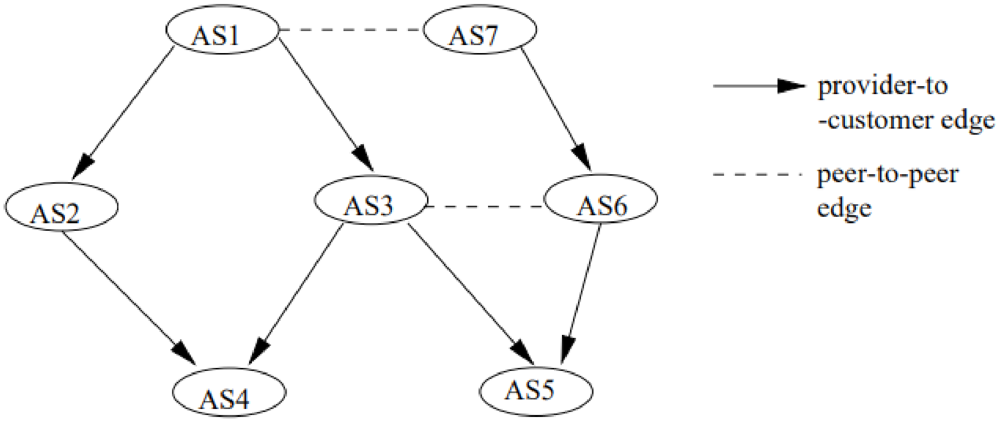
Route servers are basically the eBGP alter-egos of iBGP route reflectors. Usually they are software solutions that IXPs offer to their participants to easily interconnect to each other with a single BGP session. They run on the peering LAN of the IXP and accept BGP sessions only from BGP speakers located in the very same peering LAN of the IXP.



Picture of a route server schema, courtesy of FranceIX.net

Once a network is announced to the route server, that route is propagated to every AS connected to the route server without modifying the NEXT_HOP attribute as in every other regular BGP session. This way, the receiver will see the original NEXT_HOP attribute announced and will be able to route its packets directly via the NEXT_HOP indicated and present on the peering LAN of the IXP. Each AS connected to the route server has also the possibility to control how announced networks are re-advertised towards other peers using operational BGP communities.

In any case, AS's involved in any p2p relationship will act each other as a customer of the other. In other words, each AS will announce to the other peer only its own routes and the routes received from its customers to avoid becoming a transit for the other peer.
These economic relationships were firstly introduced by Lixin Gao in 2001 [8], and still stand for today Internet. Whenever one AS violates the above relationships with malformed or even missing filters, we have route leaks, which are defined as "the propagation of routing announcements beyond their intended scope." One of the last leaks, which caused a ripple effect across the Internet, was extensively described in one of our recent blogs.



HOW BGP ROUTING REALLY WORKS

The Internet is always in constant evolution. Nowadays there are more than 4 billion users connected to the Internet, browsing around 2 billion websites, playing games, watching videos, and doing business with each other no matter where in the world they are. This large number of users can reach their desired content via Internet routes provided by the interconnections of about 65K Autonomous Systems exchanging reachability information via BGP on about 800K different IPv4 networks and about 70K IPv6 networks. And these numbers are growing each passing minute.

The main role of BGP is "[…] to exchange network reachability information with other BGP systems." Routes are announced and withdrawn constantly from various parts of the world. Whenever a new AS joins the routing game, the first things it will do is get routes from its provider(s) to reach every Internet destination and announce to its neighbors that there is a new route towards the network(s) it owns. Each neighbor will then inform its own neighbors about these new routes, and so on so forth.

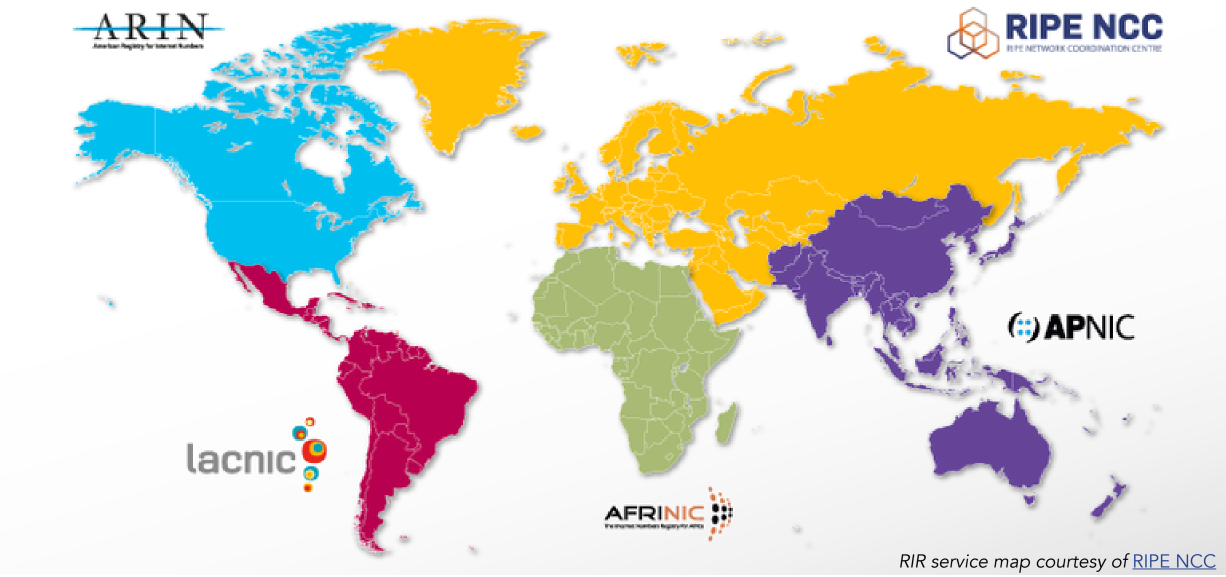
On the other hand, any AS shutting down causes the withdrawal of its routes to spread all over the Internet. But there are only a few causes of route announcement/withdrawal. A few other examples of route changes include whenever a fiber cable is accidentally cut, whenever two AS's sign a new economic agreement (or whenever that expires), and whenever there's any kind of network failure.

ROUTE ANNOUNCEMENTS AND REPLACEMENT

Once an organization gets an AS number and IPv4/IPv6 subnets from one of the five Regional Internet Registries (see the map below for the geographic distribution of RIRs) or one of the Local Internet Registries (LIRs), that organization is ready to announce its network reachability towards the global Internet.

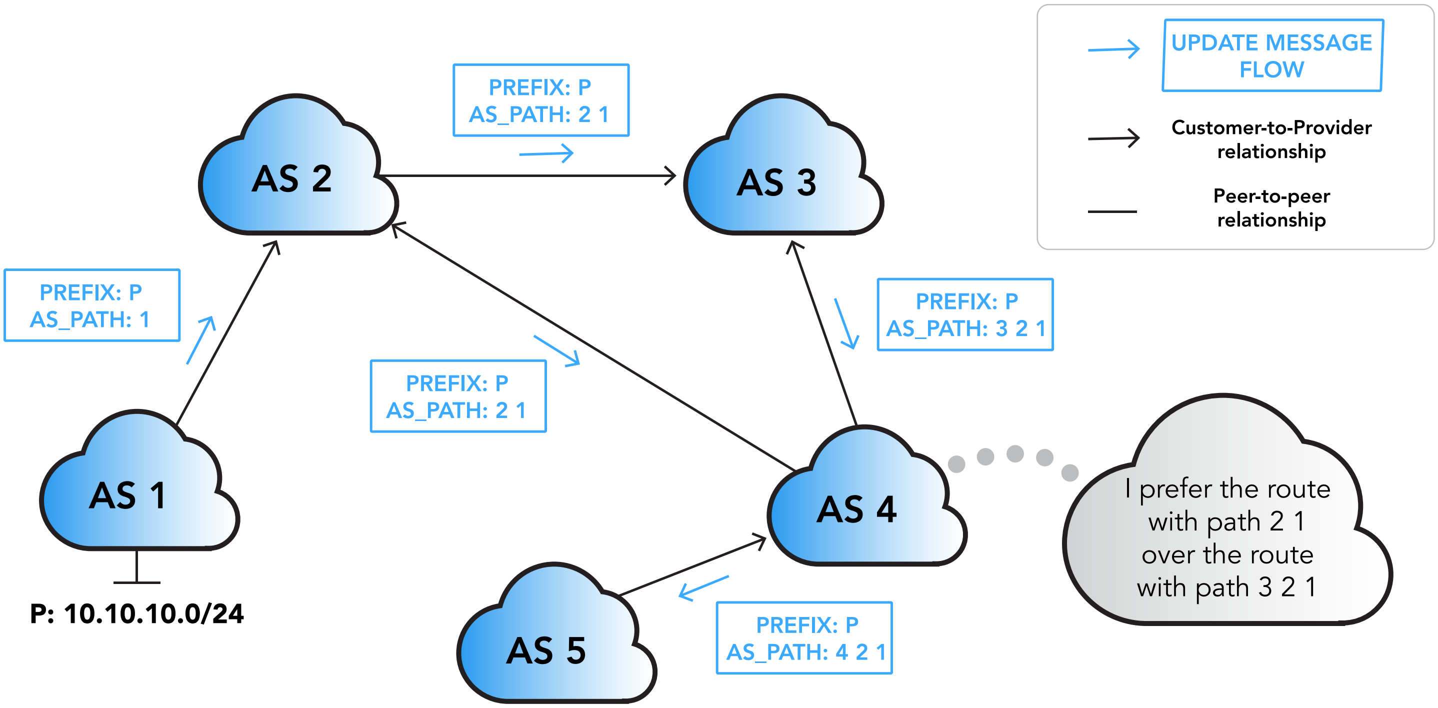
First, that organization needs to settle an agreement with one or more providers to be connected to the Internet. Then it must start advertising to the Internet the subnets obtained from the RIR/LIR so that every other player in the Internet will be aware of the presence of this new network resource and forward traffic accordingly.

To better understand how the Internet will learn about the presence of the newly advertised subnets, please consider the figure below. For ease of understanding this and the following examples, we will assume that a customer will announce only its subnets to its provider, a provider will announce everything to its customers, and a peer will announce to other peers its subnets and all the subnets it received from its customers. We will also assume that the BGP decision process will always choose as best route the route with the shortest AS path.


ARIN
American Registry for Internet Numbers
RIPE NCC
RIPE NETWORK COORDINATION CENTRE
APNIC
lacnic
AFRINIC

RIR service map courtesy of RIPE NCC

AS 1 is a brand new organization which managed to get the subnet 10.10.10.0/24 from one of the available LIRs, and which contacted AS 2 and signed an agreement so that AS 2 can transit traffic for AS 1 and allow AS 1 to be connected to the Internet. As soon as the network administrator of AS 1 installs the subnet 10.10.10.0/24 in its router, it will be generated an UPDATE message for AS 2 carrying the information that a new subnet has been announced.



More importantly, this message will carry the information that this new subnet will be reached crossing only AS 1, meaning that AS1 is the origin AS of the subnet. This information is carried in the AS path, which is one of the mandatory attributes in Update messages. This attribute is manipulated by each border router crossed to keep track of the path followed by the Update message and avoid routing loops. This is part of the BGP best route selection process.

Once AS 2 receives the Update message carrying 10.10.10.0/24, it will install this new route in its Adj-RIB-In, will select it as the best route for 10.10.10.0/24, and will install in the proper Adj-RIB-Out according to AS 2 outbound-filter policy. Thus, it will announce an Update message towards its peer, AS 3, and its other customer, AS 4, prepending its own AS number in the AS path.

At this point, AS 2 will be aware of the presence of this new subnet and will start to route traffic towards it whenever required. The same procedure will be followed by AS 3, which will propagate the Update message towards its customer AS 4, prepending its own AS number.

AS 4 will then receive two different Update messages to reach 10.10.10.0/24 at two different times. If the Update message coming from AS 3 will be received before the Update message of AS 2, then AS 5 will receive first an Update message with AS path 4 3 2 1, then another Update message with AS path 4 2 1. Otherwise, AS 5 will only receive one single Update message with AS path 4 2 1, since the piece of routing information carried by the packet announced by AS 3 will not be considered by AS 4 as best route.

Let's now assume that the BGP session between AS 2 and AS 4 is torn down. In this case, both AS's cannot reach each other anymore and the content of the related Adj-RIB-In tables will be invalidated. Consequently, the two AS's will run again the decision process for all the best routes which were involving the other AS. In the example, this means that AS 4 will analyze every other Adj-RIB-In to find a route feasible to reach 10.10.10.0/24. This process is called Path exploration and can potentially involve many routes, affecting the performances of the router. Once found a feasible replacement, then AS 4 will inform its customer that the path has changed sending an Update message with AS path 4 3 2 1.

ROUTE WITHDRAWALS

Let's now assume that the organization decides to shut down its operations. In this case, the router will be shut down and most probably sold to the highest bidder. Once the router is shut down, the BGP session established will be torn down and will trigger a domino effect all over the Internet to let everybody know that the subnets owned by the organization are no longer available to receive any traffic.

Consider once again the example shown in the figure above, with all the BGP sessions up and running. Once the BGP session between AS 1 and AS 2 is shut down, then AS 2 will start its Path exploration phase, finding no feasible routes to reach 10.10.10.0/24. Then it will generate a special Update message announcing that it cannot reach subnet 10.10.10.0/24, thus informing its neighbor to stop propagating traffic towards AS 2 to reach AS 1. AS 3 will receive this piece of information and will behave like AS 2, announcing to AS 4 that 10.10.10.0/24 is no more reachable via AS 3.

As before, AS 4 will receive two different Update messages in time. If the Update message coming from AS 3 will be received before the Update message coming from AS 2, then the first message will just remove the route from the Adj-RIB-In related to AS 3, while the second message will trigger a Path exploration phase on AS 4, which will find no feasible routes to reach 10.10.10.0/24 and will propagate the Update message to AS 5, which will be the last in line to know that the subnet has been withdrawn from the Internet.

On the other hand, if the Update message from AS 2 will be received first, then AS 4 will run a Path exploration phase which will let AS 4 believe that there is still an available route towards 10.10.10.0/24, and will advertise this new reachability to AS 5 via an Update message with AS path 4 3 2 1. In this case, only the reception of the message from AS 3 carrying the withdrawal of 10.10.10.0/24 will trigger another Path exploration phase on AS 4 and let AS 4 (and AS 5, consequently) finally understand that the route is no more there.

Please note, however, that an Update message advertising the withdrawal of a subnet does not necessarily mean that the destination is no longer reachable from any AS composing the Internet. For example, such a message could be generated in a geographic area due to a temporary local network failure and/or due to BGP session misconfigurations, while the subnet is still being reachable from other AS's.

Path exploration is a natural consequence of path vector protocols as BGP. In this family of protocols, the path information is always updated dynamically so that updates looping through the network can be discarded easily. On the other hand, the path dependencies created tend to prolong BGP protocol convergence, which can be reduced by applying special timers on the border routers.

Regional Internet Registries:

- RIPE NCC: Europe, Middle East, Russia, and parts of central Asia
- ARIN: United States, Canada, and some Caribbean countries
- APNIC: remaining part of Asia and Oceania
- LACNIC: South America, Mexico, and the remaining Caribbean countries
- AFRINIC: Africa

VULNERABILITIES OF BGP

BGP protocol has allowed network operators to apply and enforce the most varied inter-AS routing policies during the past 30 years. It is amazing how this protocol efficiently sustained the ever-increasing number of subnets and AS's, as well as the evolution of the Internet from a mostly hierarchical structure made of customers and providers to a structure where peering and IXPs become more important every day.

Despite all its good qualities, BGP shows several vulnerabilities which, if exploited, can cause ripple effects all over the Internet. The root of the problem is that BGP was conceived in an early development stage of the Internet when there were only a few players. Consequently, its design didn't consider protection against deliberate or accidental errors, so malicious or misconfigured sources can potentially propagate fake routing information all over the Internet, exploiting this lack of protection. Even worse, the source of fake or malicious routing information could be either a real BGP peer or a fake peer, since BGP runs on TCP/IP and is consequently subject to every classic TCP/IP attack such as IP spoofing.

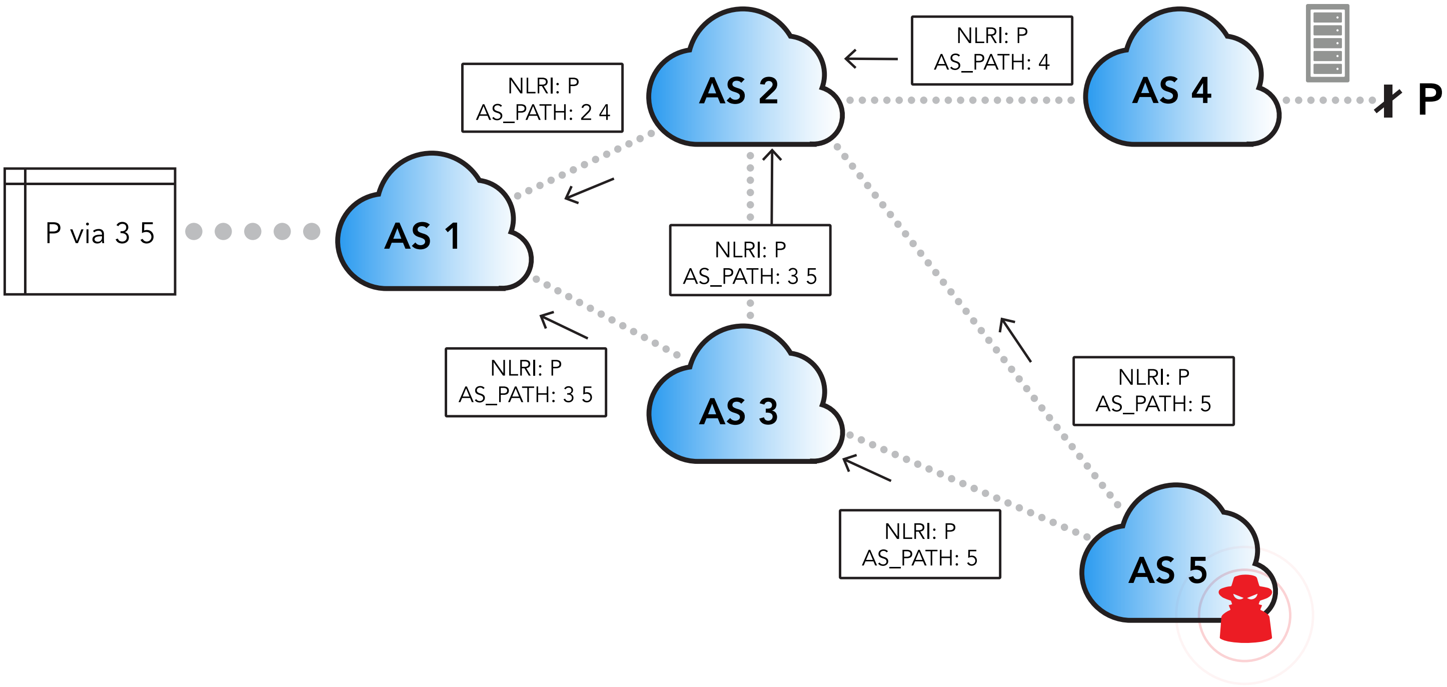
Part of the problem can be solved applying cryptographic authentication on each BGP peer, but this won't help stop bogus information spreading all over the Internet from legitimate misconfigured sources (route leaks), from legitimate sources which either didn't apply cryptographic authentication at all, or from sources that deliberately announced bogus routing information (prefix hijacks).

Solutions like Resource Public Key Infrastructure (RPKI) and BGPsec path validation have been recently standardized by IETF, but they still require the collaboration of many AS's and thus are difficult to deploy.

PREFIX HIJACK ATTACKS

Prefix hijacks are deliberate intentional generation of bogus routing information; the reasons behind them are of a multitude that is difficult to fathom.

The attacker could announce routes to disrupt the services running on top of the IP space covered by the routes, or hijack the traffic to analyze confidential information flowing towards that service. The attacker could also simply announce routes with a crafted AS path to show fake neighboring connections in famous websites, like the BGP toolkit of Hurricane Electric. Or even worse, the attacker could hijack the traffic to manipulate the flowing packets at his/her will, or simply want to exploit unused routes to generate spam.

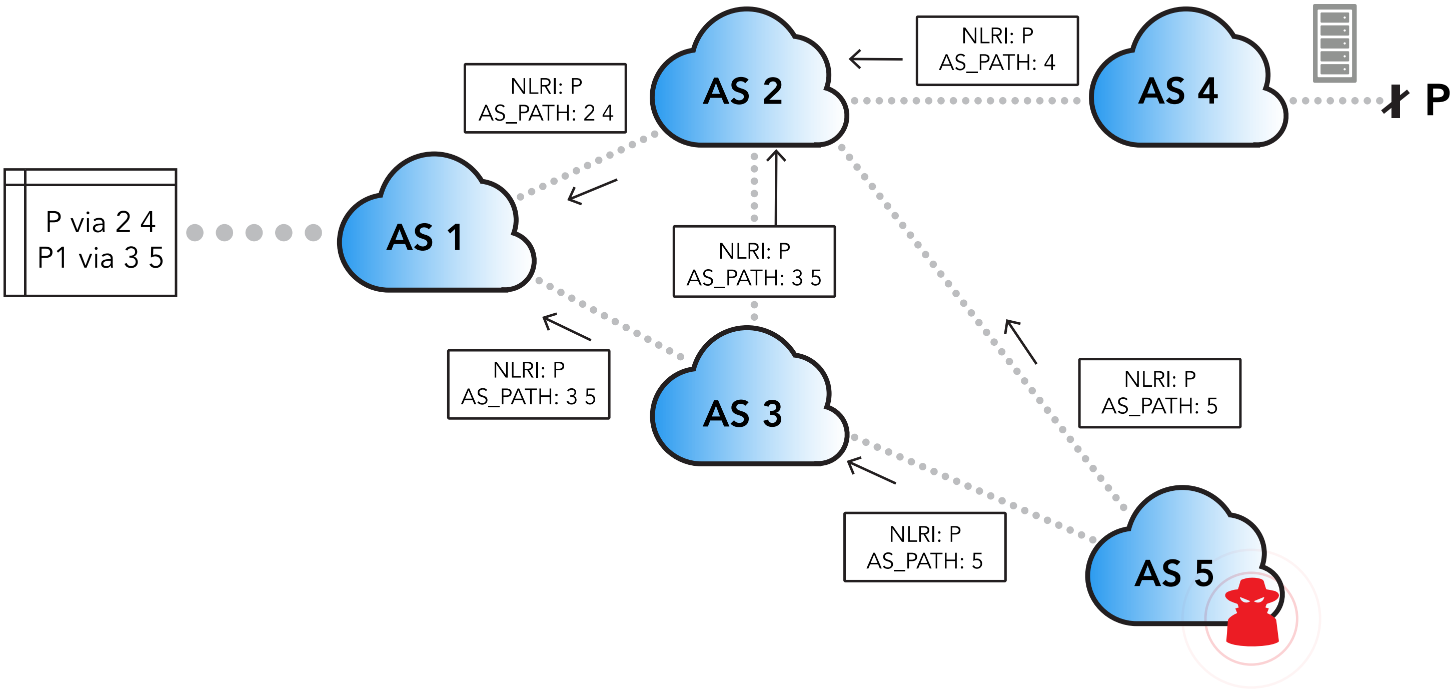


Let's consider the above scenario to better understand how prefix hijacks can be performed; we will consider the following topology in this and in the following examples. AS 5 is a malicious attacker and is connected to the Internet via two providers: AS 2 and AS 3. AS 1 is customer of AS 2 and provider of AS 3, while AS 4 is a peer of AS 2 and AS 2 is provider of AS 3. Finally, we assume that AS 2 has properly set its incoming BGP filters, while AS 1 and AS 3 have a loose filter configuration (if any).

In this first scenario, AS 5 will announce network P, which is owned and already announced by AS 4. Due to the filter configurations described above, the Update message announced by AS 5 will be dropped by AS 2, while it will be accepted by AS 3. AS 3 will then announce that to its providers (AS 1 and AS 2). AS 2 will again drop the packet due to the filters, while AS 1 will accept it. If the BGP decision process of AS 1 will select as best route the path from AS 5, then traffic from AS 1 to AS 5 will be sent to the attacker instead of towards the proper owner.

Consider now this example to be composed of about 60,000 AS's, each with its own filter policy, if any. The consequence is that part of the Internet will redirect its traffic towards the attacker, while the rest will redirect its traffic towards the proper origin. The amount of AS's redirecting their traffic towards the attacker will depend on two factors: the quality of the filters applied by the providers, and the BGP decision process output of each AS.

Note that in this scenario it is possible to identify the attacker by checking BGP packets involving P either from route collectors (with a proper post-mortem analysis), via dedicated real-time BGP monitoring systems, and via customer complaints, since traffic is not re-directed to the original owner.



Let’s now consider another scenario on the very same topology. AS 5 will now announce network P1 subnet of network P, still owned by AS 4 but never advertised by AS 4. For example, consider P to be 10.0.0.0/23, then P1 could either be 10.0.0.0/24 or 10.0.1.0/24. AS 5 will announce it only to AS 3, knowing that AS 3 filters are loose. In addition, AS 5 will know that AS 2’s filters are tight and will exploit that to keep a safe route towards the destination.

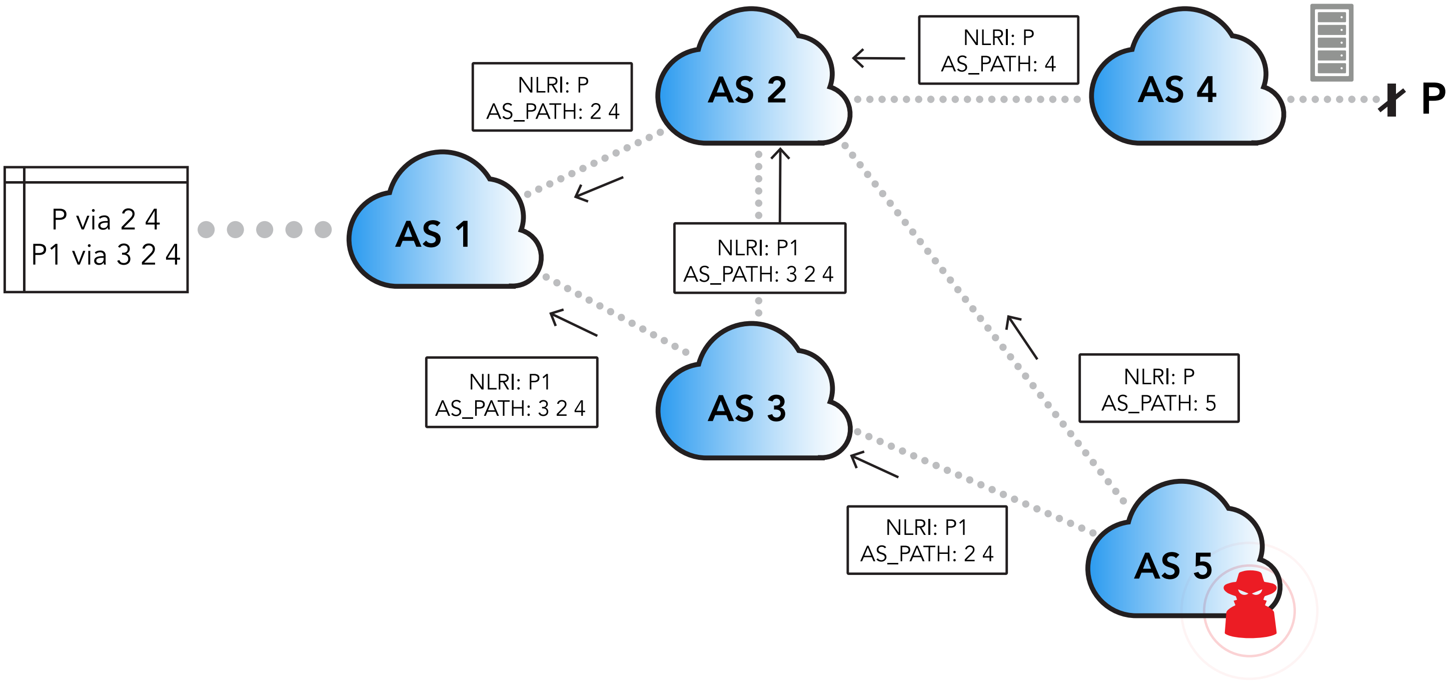
In this scenario, P1 will propagate the same way as P in the previous scenario. The slight difference is that now every affected AS will have two different routes for the IP space covered by P: P and P1. Let’s focus on AS 1.

Even if a proper route to P is installed in AS 1’s router, only a portion of traffic of the original P will be directed to the proper owner due to the longest prefix match. Please note that since AS 5 kept one of its providers explicitly out of the hijack, AS 5 can now route traffic received from AS 1 directed to P1 to the proper owner, after analyzing and/or manipulating each packet.

Now consider again this real-world example and imagine that AS 4 is hosting on P1 some servers of a bank. Consider now that the attacker is interested in collecting data from the bank, and that he/she studied the problem deeply enough to know that P1 is the ideal target for its purposes and starts announcing it. Differently from the previous scenario, the bogus routing information spread will now depend only on the quality of the filters applied by AS’s, since the subnet P1 and P will not interfere with each other in BGP decision processes. As soon as everything is set up, then AS 5 will be able to receive data from the affected portion of the world, while keeping a safe routing leg to forward traffic and (hopefully for him/her) get unnoticed.

Again, note that in this scenario it is still possible to identify the attacker by checking BGP packets involving P and any subnet of P either from route collectors or via dedicated real-time BGP monitoring systems. However, the network operator can’t identify the attack from the complaints received by his/her customers if the delay introduced by the attacker is short enough to go unnoticed.

An example of a route leak which falls perfectly in this scenario is the infamous hijack of YouTube prefixes by Pakistan Telecom back in late February 2008. In that case, Pakistan Telecom attempted to blackhole traffic towards 208.65.153.0/24 by announcing routes where Pakistan Telecom was appearing as the origin AS to fulfill a censorship request from the Pakistan government. The problem is that they also announced this route to its provider PCCW, which didn't apply proper filters and caused a domino effect, causing about 3 hours of service disruption to YouTube.

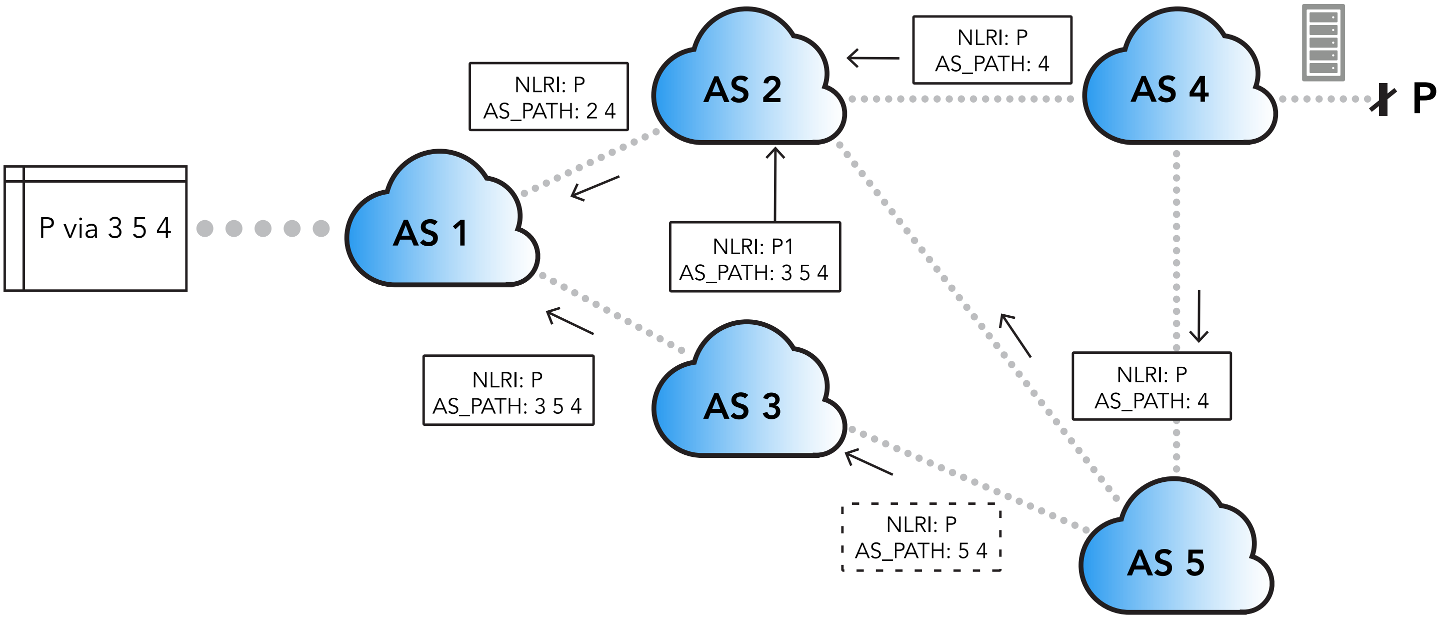


Consider now the above scenario. AS 5 is now smart enough to forge a fake AS path in the Update message by keeping the AS of the real owner at the end of the AS path as well as the original provider of the real owner (AS 2).

The propagation of the attack is the same as the previous examples, but now the detection of the attack is much harder. It is still possible to check BGP packets involving P and any subnet of P either from route collectors or via dedicated real-time BGP monitoring systems, but now the detection of the attack must also rely on additional pieces of information, such as the knowledge of each relationship between each pair of AS's in the AS path. Indeed, in this example it would have been possible to detect that since AS 3 is customer of AS 2, and the AS path 3 2 4 detected at AS 1 would have shown the involvement of AS 3 as transit of AS 2 for P1, which is against the valley-free property.

ROUTE LEAKS AND FAT FINGER SYNDROME

Route leaks are unintentional generation of bogus routing information caused by router misconfigurations, such as typos in the filter configuration or mis-origination of someone's else network (fat finger). Even if unintentional, the consequences of a route leak can be the same as the prefix hijacks.



Consider the very same topology we used in the prefix hijack examples, with the difference that AS 5 is now a normal network operator which simply applied wrong BGP filters, such as "accept everything from my provider, announce everything to my provider." This is sadly not an uncommon case, and it is an error that several AS's can do when switching from a single provider (where this rule works fine) to multiple providers (where this rule would make the AS a transit of each provider).

Due to that mistake, now AS 5 will propagate everything it receive from its provider towards another provider, clearly against the valley-free property. This piece of routing information will then spread all over the Internet and AS's will start routing traffic depending on the result of the BGP decision process of each AS.

Now think again about the 65,000 AS's in the Internet and imagine that AS 4 is a rural service provider with few resources, both technical and economic. This would mean that probably the upstream connection he/she bought from his/her providers is very limited, thus making the two links a bottleneck in this route leak scenario. In this case it is possible that AS 5 will not be able to handle the amount of traffic directed to *P*, causing not only an additional delay, but also several packet losses.

This was the case of the route leak we discussed in our June blog, which affected several banks in addition to Facebook and CloudFlare. This wasn't the only case of route leak recently experienced, and thanks to that IETF managed to draw a remarkable route leak classification.

catchpoint™

A DIFFERENT APPROACH TO DIGITAL EXPERIENCE MONITORING

Catchpoint, the global leader in Digital Experience Monitoring (DEM), empowers business and IT leaders to protect and advance the experience of their customers and employees. In a digital economy, enabled by cloud, SaaS and IoT, applications and users are everywhere. Catchpoint offers the largest and most geographically distributed monitoring network in the industry – it's the only DEM platform that can scale and support today's customer and employee location diversity and application distribution. It helps enterprises proactively detect, identify, and validate user and application reachability, availability, performance and reliability, across an increasingly complex digital delivery chain. Industry leaders like Google, L'Oréal, Verizon, Oracle, LinkedIn, Honeywell, and Priceline trust Catchpoint's out-of-the box monitoring platform, to proactively detect, repair, and optimize customer and employee experiences.

To request a free trial, visit www.catchpoint.com/trial