Anatomy of a route leak

Jérôme Fleury - Director of Network Engineering
Tom Strickx - Network Software Engineer
Martin J Levy - Distinguished Engineer
Introduction
Anatomy of this talk

In the following thirty minutes ...

• Some Internet history
• Some BGP route leak history
• Something happened June 2019
• Some route optimizer comments
• Some graphs
March 1977 - no routing security

The Internet was not built for what it has become
Security

This option provides a way for hosts to send security, compartmentalization, handling restrictions, and TC (closed user group) parameters. The format for this option is as follows:

```
+--------------------------------+---------+---------+---------+
| 10000010 00001011 SSS SSS CCC CCC CCC HHH HHH HHH TCC |
+--------------------------------+---------+---------+---------+
Type=130 Length=11
```

Security (S field): 16 bits

Specifies one of 16 levels of security (eight of which are reserved for future use).

<table>
<thead>
<tr>
<th>Security Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000000 00000000</td>
<td>Unclassified</td>
</tr>
<tr>
<td>11110001 00110101</td>
<td>Confidential</td>
</tr>
<tr>
<td>01111000 10011010</td>
<td>EFTO</td>
</tr>
<tr>
<td>10111100 01001101</td>
<td>MMMM</td>
</tr>
<tr>
<td>01011110 00100110</td>
<td>PROG</td>
</tr>
<tr>
<td>10101111 00010111</td>
<td>Restricted</td>
</tr>
<tr>
<td>11010111 10001000</td>
<td>Secret</td>
</tr>
<tr>
<td>01101011 11000101</td>
<td>Top Secret</td>
</tr>
<tr>
<td>00110101 11100010</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>10011101 11110001</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>01001101 01111000</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>00100100 10111101</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>00101011 01011110</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>10001001 10101111</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>11000100 11010110</td>
<td>(Reserved for future use)</td>
</tr>
<tr>
<td>11100010 01101011</td>
<td>(Reserved for future use)</td>
</tr>
</tbody>
</table>

Compartment (C field): 16 bits

An all zero value is used when the information transmitted is not compartmented. Other values for the compartments field may be obtained from the Defense Intelligence Agency.

Handling Restrictions (R field): 16 bits

The values for the control and release markings are alphanumeric digraphs and are defined in the Defense Intelligence Agency Manual DIAM 65-19, "Standard Security Markings".
The World Wide Web comes from CERN (Geneva Switzerland)

1989/1990 CERN

Information Management: A Proposal

Tim Berners-Lee, CERN
March 1989, May 1990

Non requirements

Discussions on Hypertext have sometimes tackled the problem of copyright enforcement and data security. These are of secondary importance at CERN, where information exchange is still more important than secrecy. Authorisation and accounting systems for hypertext could conceivably be designed which are very sophisticated, but they are not proposed here.

In cases where reference must be made to data which is in fact protected, existing file protection systems should be sufficient.
1991 RFC1267 - BGP3

RFC793 is the first definition of TCP

2.9. Precedence and Security

Security Considerations

Security issues are not discussed in this memo.
Insecure yesterday, Secure today

Yesterday:

Connection ← Name ← IP ← ASN

DNS ↑ ↓  RIRs ↑ ↓

Today:

Connection ← PKI cert → Name ← IP ← ASN

DNS ↑ ↓  → RIRs ↑ ↓

(~200 million)  (~1 billion)  (~700K routes)

We verify

We encrypt
How it looks to the press

Amazon, Facebook internet outage: Verizon blamed for 'cascading catastrophic failure'

DHS issues security alert about recent DNS hijacking attacks

Google traffic hijacked via tiny Nigerian ISP
BGP’s timeline of leaks

1997: The "AS 7007 incident"
- Triggered by a router bug

2008: Pakistan Telecom hijacks YouTube
- Regulatory hijack

2015: Malaysia Telecom
- Propagated by Level3

2017: Google leaks to Verizon
- Starts in Chicago, impact Japan

2018: MainOne leaks Google, Cloudflare
- Leaked to China Telecom, raises suspicions

2019: Verizon leak
- BGP optimizers make it worse
February 5\textsuperscript{th}, 2020 ... the beat goes on!

It would seem that it’s not safe to run a “\textit{quad address}”!

https://twitter.com/DougMadory/status/1225486918609227777
June 24th, 2019, 10:30 UTC
June 24th, 2019 - 10:30 UTC

Cloudflare issues affecting numerous sites on Monday AM [Update: fixed]

- Slate: An internet outage caused by DQE and apparently Verizon shows how fragile the web is.
- The Associated Press: Cloudflare Chief Technology Officer John Graham-Cumming told the Washington Post that Verizon failed to intercept the issue from a fiber-optic network services provider.

Customers report Verizon, Cloudflare disruptions.
June 24th, 2019 leak was widespread

Source: Cedexis
June 24th, 2019 Impact on the Cloudflare traffic
How did it get solved?
Problem Definition and Classification of BGP Route Leaks

Abstract

A systemic vulnerability of the Border Gateway Protocol routing system, known as "route leaks", has received significant attention in recent years. Frequent incidents that result in significant disruptions to Internet routing are labeled route leaks, but to date a common definition of the term has been lacking. This document provides a working definition of route leaks while keeping in mind the real occurrences that have received significant attention.

Further, this document attempts to enumerate (though not exhaustively) different types of route leaks based on observed events on the Internet. The aim is to provide a taxonomy that covers several forms of
A very invalid route - step #1

```
104.20.56.0/21  unicast [nforce1_4 10:34:29.282] * (100) [AS13335?]
   via 185.107.95.164 on eno1
Type: BGP univ ,-- "Allegheny Technologies Incorporated"
BGP.origin: Incomplete  v
BGP.as_path: 43350 6762 701 396531 33154 3356 13335
BGP.next_hop: 185.107.95.164
BGP.local_pref: 100

   unicast [nforce2_4 10:34:29.296] (100) [AS13335?]
   via 185.107.95.165 on eno1
Type: BGP univ
BGP.origin: Incomplete
BGP.as_path: 43350 6762 701 396531 33154 3356 13335
BGP.next_hop: 185.107.95.165
BGP.local_pref: 100
```
A very invalid route - step #2

<table>
<thead>
<tr>
<th>Prefix</th>
<th>104.25.48.0/20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Length</td>
<td>/20</td>
</tr>
<tr>
<td>ASN</td>
<td>13335</td>
</tr>
<tr>
<td>Trust Anchor</td>
<td>ARIN</td>
</tr>
<tr>
<td>Validity</td>
<td>Thu, 02 Aug 2018 04:00:00 GMT - Sat, 31 Jul 2027 04:00:00 GMT</td>
</tr>
<tr>
<td>Emitted</td>
<td>Thu, 02 Aug 2018 21:45:37 GMT</td>
</tr>
<tr>
<td>Name</td>
<td>535ad55d-dd30-40f9-8434-c17fc413aa99</td>
</tr>
<tr>
<td>Key</td>
<td>4a75b5de16143adbeaa987d6d91e0519106d086e</td>
</tr>
<tr>
<td>Parent Key</td>
<td>a6e7a6b44019cf4e388766d940677599d0c492dc</td>
</tr>
<tr>
<td>Path</td>
<td>rsync://rpki.arin.net/repository/arin-rpki-ta/5e4a23ea-...</td>
</tr>
</tbody>
</table>
The disruptive power of Tier 1 providers
We wrote two blogs about all this

How Verizon and a BGP Optimizer Knocked Large Parts of the Internet Offline Today

June 24, 2019 12:58 PM

Massive route leak impacts major parts of the Internet, including Cloudflare

What happened?

Today at 10:30UTC, the Internet had a small heart attack. A small company in Northern Pennsylvania became a preferred path of many Internet routes through Verizon (AS701), a major Internet transit provider. This was the equivalent of Waze routing an entire freeway down a neighborhood street — resulting in many websites on Cloudflare, and many other providers, to be unavailable from large parts of the Internet. This should never have happened because Verizon should never have forwarded those routes to the rest of the Internet. To understand why, read on.


The deep-dive into how Verizon and a BGP Optimizer Knocked Large Parts of the Internet Offline Monday

June 26, 2019 3:22 PM

A recap on what happened Monday

On Monday we wrote about a painful Internet wide route leak. We wrote that this should never have happened because Verizon should never have forwarded those routes to the rest of the Internet. That blog entry came out around 19:58 UTC, just over seven hours after the route leak finished (which will we see below was around 12:39 UTC). Today we will dive into the archived routing data and analyze it. The format of the code below is meant to use simple shell commands so that any reader can follow along and, more importantly, do their own investigations on the routing tables.

This was a very public BGP route leak event. It was both reported online via many news outlets and the event’s BGP data was reported via social media as it was happening. Andree Toonk tweet a quick list of 2,400 ASNs that were affected.

We included all the scripting to show leaks

```bash
$ # Collect 24 hours of data - more than enough
$ ASN="AS13335"
$ START="2019-06-24T00:00:00"
$ END="2019-06-25T00:00:00"
$ ARGS="resource=$ASN&starttime=$START&endtime=$END"
$ URL="https://stat.ripe.net/data/bgp-updates/data.json?$ARGS"
$ # Fetch the data from RIPEstat
$ curl -sS "$URL" | jq . > 13335-routes.json
$ ls -l 13335-routes.json

$ # Extract just the times, routes, and AS-PATH
$ jq -rc '.data.updates[].timestamp,.attrs.target_prefix,.attrs.path' < 13335-routes.json | paste - - - > 13335-listing-a.txt
$ wc -l 13335-listing-a.txt
691318

$ # Extract the route leak 701,396531
$ # AS701 is Verizon and AS396531 is Allegheny Technologies
$ egrep '701,396531' < 13335-listing-a.txt > 13335-listing-b.txt
$ wc -l 13335-listing-b.txt
204568

$ # Extract the actual routes affected by the route leak
$ cut -f2 < 13335-listing-b.txt | sort -V -u > 13335-listing-c.txt
$ wc -l 13335-listing-c.txt
101
```

So far this is IPv4 speak - what about IPv6?

IPv6? Where is the IPv6 route leak?

In what could be considered the only plus from Monday’s route leak, we can confirm that there was no route leak within IPv6 space. Why?

It turns out that 396531 (Allegheny Technologies Inc) is a network without IPv6 enabled. Normally you would hear Cloudflare chastise anyone that’s yet to enable IPv6, however, in this case we are quite happy that one of the two protocol families survived. IPv6 was stable during this route leak, which now can be called an IPv4-only route leak.

Yet that’s not really the whole story. Let’s look at the percentage of traffic Cloudflare sends Verizon that’s IPv6 (vs IPv4). Normally the IPv4/IPv6 percentage holds steady.
Peerlock
Peerlock

Ideal for (tier1) transit networks: reject any route from your customers that contains another “big boy” in the AS Path:

174_701_396531_33154_3356_13335

If you’re Cogent (AS174), you have no reason to accept this route from Verizon (AS701) that contains Level3 (AS3356) within the path.

Even if you’re not a Tier1, you can apply this to your customers sessions!

All tier1's have direct interconnection with other tier1's. Financial relationships are not diagramed, this is only routing.
Peerlock - easier for Tier1’s vs others

The absolute definition of a Tier1 makes their job easier

Content networks towards transits - doable

IXP filterings - much harder (but worthy of thought)
BGP route optimizers
BGP route optimizers make it worse

So-called “BGP optimizers” use a technique that deaggregate existing BGP routes into smaller prefixes so that your router can load-balance traffic over multiple links.

If you leak these “fake” routes, you will attract all Internet traffic for these... unless your upstreams filter them.
BGP optimizers to make it worse

This is an invalid route
BGP optimizers - our view

Do route optimizers cause fake routes?

Mar 27, 2015

Yesterday, an incident occurred involving Autonomous System 65535 (AS65535). We’re confident that the AS65535 network, which is “more specific” than the AS belonging to the CloudFlare networks, was configured to act as a “false root.”

CloudFlare networks, which are intended to be part of the AS, were mistakenly advertised to the AS65535 network. As a result, traffic should have been filtered to CloudFlare networks, but instead was redirected elsewhere, leading to suboptimal routing and degraded performance for customers of CloudFlare.

According to our analysis, the AS65535 network did not have the correct policy in place to filter these routes correctly. This resulted in traffic going to an incorrect AS, which caused performance issues for CloudFlare customers.

BGP Filtering Best Practices

This eBook discusses BGP Filtering and provides configurations needed to set up filters with public and private peers, upstream providers as well as downstream customers.
BGP optimizer - leaking by default

In order to further reduce the likelihood of these problems occurring in the future, we will be adding a feature within Noction IRP to give an option to tag all the more specific prefixes that it generates with the BGP NO_EXPORT community. This will not be enabled by default, due to potential drawbacks; such as customers who use multiple ASes or customers who have eBGP sessions with private ASes, but it will be an option if a customer wants to use it. This way, even if filters fail, more specific prefixes won't be propagated to external autonomous systems.

... option to tag all the more specific prefixes that it generates with the BGP NO_EXPORT community. **This will not be enabled by default**
Noction response

Noction responds regarding June/24 route leak.

In fact, the use of more specific prefixes is only going to increase no matter if a network uses any BGP tools or not. In this specific case, the more specific prefixes were generated by Noction IRP.

[...]

Unfortunately, BGP is not perfect. Almost 2300 leaks or hijacks happened over the past 7 months. Poor use of filters at Tier 1, Tier 2 and Tier 3 levels linked to all of them.

[...]

NO_EXPORT is not a good option for companies operating multiple ASNs, be it multiple public or a combination of private and public.
RPKI (because ROA is required)
What can we do about it?

- Apply best practices:
  - MANRS - [https://www.manrs.org/](https://www.manrs.org/)
- IRR filtering is easier said than done.
  - There is no recipe to build prefix filters and a lot of questions remain unanswered:
    - How often should you update your prefix filters?
    - What IRR database should you trust?
    - What automation framework should you use?
    - How do you deliver feedback to your peers?
2018-2020 are big years for Routing Security

- Cloudflare issued route origin authorizations ("ROAs")
  - covers 90+% of its prefixes, including:
    - Its 1.1.1.1 resolver
    - DNS servers
- NTT now treats ROAs as if they were IRR route(6)-objects
- **AS7018/AT&T, AS286/KPN AS1299/Telia now dropping RPKI invalids**
- 100+ networks have joined the Mutually Agreed Norms for Routing Security ("MANRS")
- Google to begin filtering routes in 2019
- **ARIN allowed integration of its contract into RPKI software workflows and renewed its review of legal issues**
Hello all.

In November 2018 during the ZAPF (South Africa Peering Forum) meeting in Cape Town, 3 major ISP’s in Africa announced that they would enable RPKI’s ROV (Route Origin Validation) and the dropping of Invalid routes as part of an effort to clean up the BGP Internet, on the 1st April, 2019.

On the 1st of April, Workonline Communications (AS37271) enabled ROV and the dropping of Invalid routes. This applies to all eBGP sessions for IPv4 and IPv6.

On the 5th of April, SEACOM (AS37100) enabled ROV and the dropping of Invalid routes. This applies to all eBGP sessions with public peers, private peers and transit providers, both for IPv4 and IPv6. eBGP sessions toward downstream customers will follow in 3 months from now.

We are still standing by for the 3rd ISP to complete their implementation, and we are certain they will communicate with the community accordingly.

Please note that for the legal reasons previously discussed on various fora, neither Workonline Communications nor SEACOM are utilising the ARIN TAL. As a result, any routes covered only by a ROA issued under the ARIN TAL will fall back to a status of Not Found. Unfortunately, this means that ARIN members will not see any improved routing security for their prefixes on our networks until this is resolved. We will each re-evaluate this decision if and when ARIN’s policy changes. We are hopeful that this will happen sooner rather than later.

AS37100/Seacom does not use the ARIN TAL; hence routes allocated by ARIN were not protected.
Lowering Legal Barriers to RPKI Adoption

Deploy RPKI now (Because tomorrow is already too late)
AS7018/AT&T AS1299/Telia and RPKI

Job Snijders
@JobSnijders

BREAKING - AT&T / AS 7018 is now rejecting RPKI Invalid BGP announcements they receive from their peering partners. This is big news for routing security! If AT&T can do it - you can do it! :-)

mailman.nanog.org/pipermail/nano...

472  6:09 PM - Feb 11, 2019

248 people are talking about this

Telia Carrier
@TeliaCarrier

Telia Carrier/AS1299 is now as the first Tier-1 dropping RPKI invalid prefixes from both customers & peers. 😊

#RPKI

8:53 AM - Feb 5, 2020 - Sprout Social

24 Retweets  48 Likes
Summary
Questions?

martin @cloudflare.com
jf   @cloudflare.com
tstrickx @cloudflare.com
Additional content
Security was always being discussed and defined; but mainly in computing.
RFC793 is the first definition of TCP

2.9. Precedence and Security

The TCP makes use of the internet protocol type of service field and security option to provide precedence and security on a per connection basis to TCP users. Not all TCP modules will necessarily function in a multilevel secure environment; some may be limited to unclassified use only, and others may operate at only one security level and compartment. Consequently, some TCP implementations and services to users may be limited to a subset of the multilevel secure case.

TCP modules which operate in a multilevel secure environment must properly mark outgoing segments with the security, compartment, and precedence. Such TCP modules must also provide to their users or higher level protocols such as Telnet or TFTP an interface to allow them to specify the desired security level, compartment, and precedence of connections.
Without proper filtering
With proper filtering
Border Gateway Protocol (BGP) picks the best routes for data to travel, which usually means hopping between autonomous systems.

Each Autonomous System uses BGP routing to send packets between systems until they reach their destination.

More interconnection = more opportunity to share route information.
BGP’s sad timeline of leaks

A small subset of BGP’s global route leaks:

- April 1997: The "AS 7007 incident"
- February 2008: YouTube globally routed into Pakistan Telecom
- April 2010: Chinese ISP hijacks the internet
- April 2014: Indosat leaks
- June 2015: Malaysia Telecom
- August 2017: Google leaks to Verizon
- November 2018: MainOne leaks Google, Cloudflare
- June 2019: Verizon leaks