Anatomy of a route leak

Jérôme Fleury - Director of Network Engineering
Tom Strickx - Network Software Engineer
Martin J Levy - Distinguished Engineer

Introduction

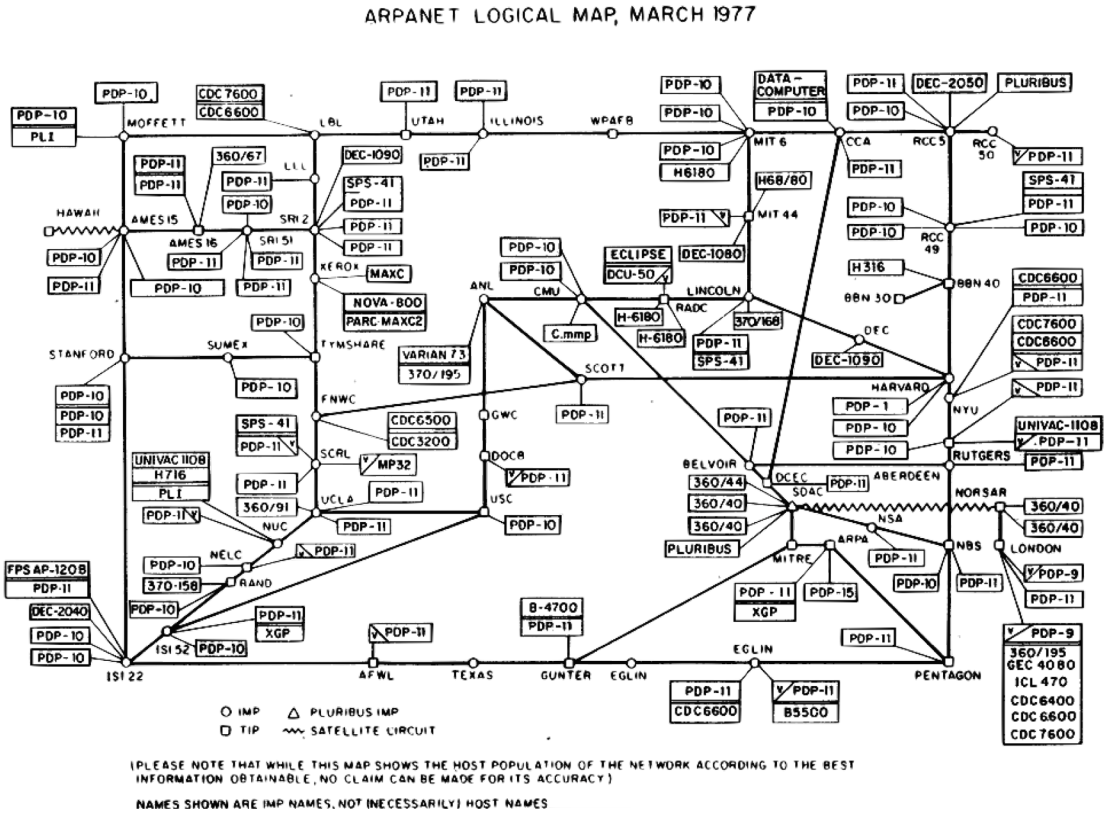
Anatomy of this talk

In the following thirty minutes ...

- Some Internet history
- Some BGP route leak history
- Something happened June 2019
- Some route optimizer comments
- Some graphs

CLOUDFLARE

March 1977 - no routing security



The Internet was not built for what it has become

CLOUDFLARE

1981

Security

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

```
+--------+--------+---//---+---//---+---//---+---//---+
|10000010|00001011|SSS  SSS|CCC  CCC|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).

```
00000000 00000000 - Unclassified
11110001 00110101 - Confidential
01111000 10011010 - EFTO
10111100 01001101 - MMMM
01011110 00100110 - PROG
10101111 00010011 - Restricted
11010111 10001000 - Secret
01101011 11000101 - Top Secret
00110101 11100010 - (Reserved for future use)
10011010 11110001 - (Reserved for future use)
01001101 01111000 - (Reserved for future use)
00100100 10111101 - (Reserved for future use)
00010011 01011110 - (Reserved for future use)
10001001 10101111 - (Reserved for future use)
11000100 11010110 - (Reserved for future use)
11100010 01101011 - (Reserved for future use)
```


Compartments (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 (H 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".

CLOUDFLARE

RFC791 is the first definition of IP

Section 3.1. Internet Header Format

Security option type=130

1989/1990 CERN

Information Management: A Proposal

Tim Berners-Lee, CERN

March 1989, May 1990

Non requirements

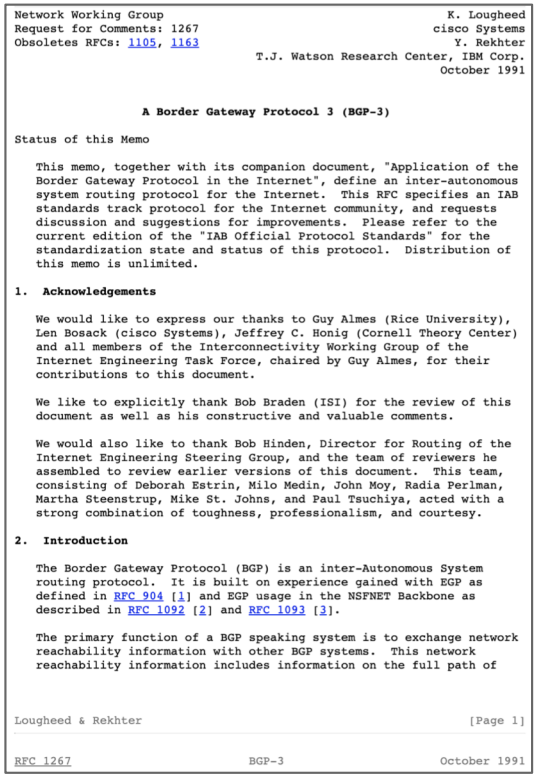
Discussions on Hypertext have sometimes tackled the problem of copyright enforcement and data security. These are of secondary importance at CERN, where information exchange is still more important than secrecy. Authorisation and accounting systems for hypertext could conceivably be designed which are 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.

The World Wide Web comes from CERN (Geneva Switzerland)

CLOUDFLARE

1991 RFC1267 - BGP3



Network Working Group
Request for Comments: 1267
Obsoletes RFCs: 1105, 1163

K. Lougheed
cisco Systems
Y. Rekhter
T.J. Watson Research Center, IBM Corp.
October 1991

A Border Gateway Protocol 3 (BGP-3)

Status of this Memo

This memo, together with its companion document, "Application of the Border Gateway Protocol in the Internet", define an inter-autonomous system routing protocol for the Internet. This RFC specifies an IAB standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "IAB Official Protocol Standards" for the standardization state and status of this protocol. Distribution of this memo is unlimited.

1. Acknowledgements

We would like to express our thanks to Guy Almes (Rice University), Len Bosack (cisco Systems), Jeffrey C. Honig (Cornell Theory Center) and all members of the Interconnectivity Working Group of the Internet Engineering Task Force, chaired by Guy Almes, for their contributions to this document.

We like to explicitly thank Bob Braden (ISI) for the review of this document as well as his constructive and valuable comments.

We would also like to thank Bob Hinden, Director for Routing of the Internet Engineering Steering Group, and the team of reviewers he assembled to review earlier versions of this document. This team, consisting of Deborah Estrin, Milo Medin, John Moy, Radia Perlman, Martha Steenstrup, Mike St. Johns, and Paul Tsuchiya, acted with a strong combination of toughness, professionalism, and courtesy.

2. Introduction

The Border Gateway Protocol (BGP) is an inter-Autonomous System routing protocol. It is built on experience gained with EGP as defined in RFC 904 [1] and EGP usage in the NSFNET Backbone as described in RFC 1092 [2] and RFC 1093 [3].

The primary function of a BGP speaking system is to exchange network reachability information with other BGP systems. This network reachability information includes information on the full path of

Lougheed & Rekhter [Page 1]

RFC 1267 BGP-3 October 1991

Security Considerations

Security issues are not discussed in this memo.

Security issues are not discussed in this memo.

RFC793 is the first definition of TCP

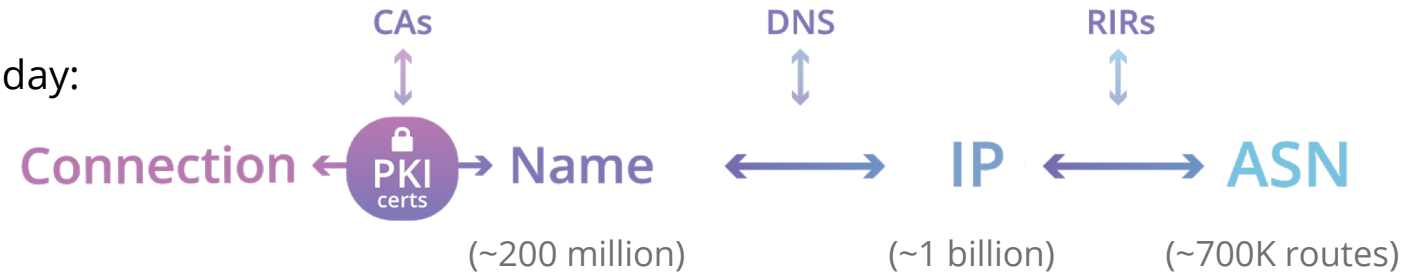
2.9. Precedence and Security

CLOUDFLARE

Insecure yesterday, Secure today

Yesterday:

Today:

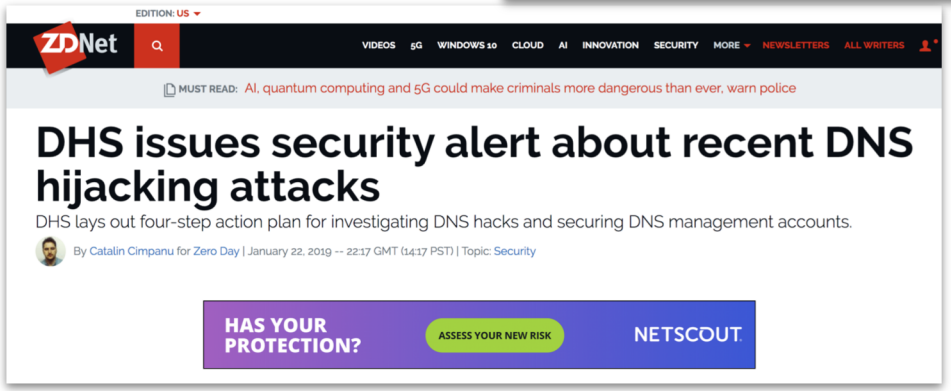


CLOUDFLARE

We verify

We encrypt

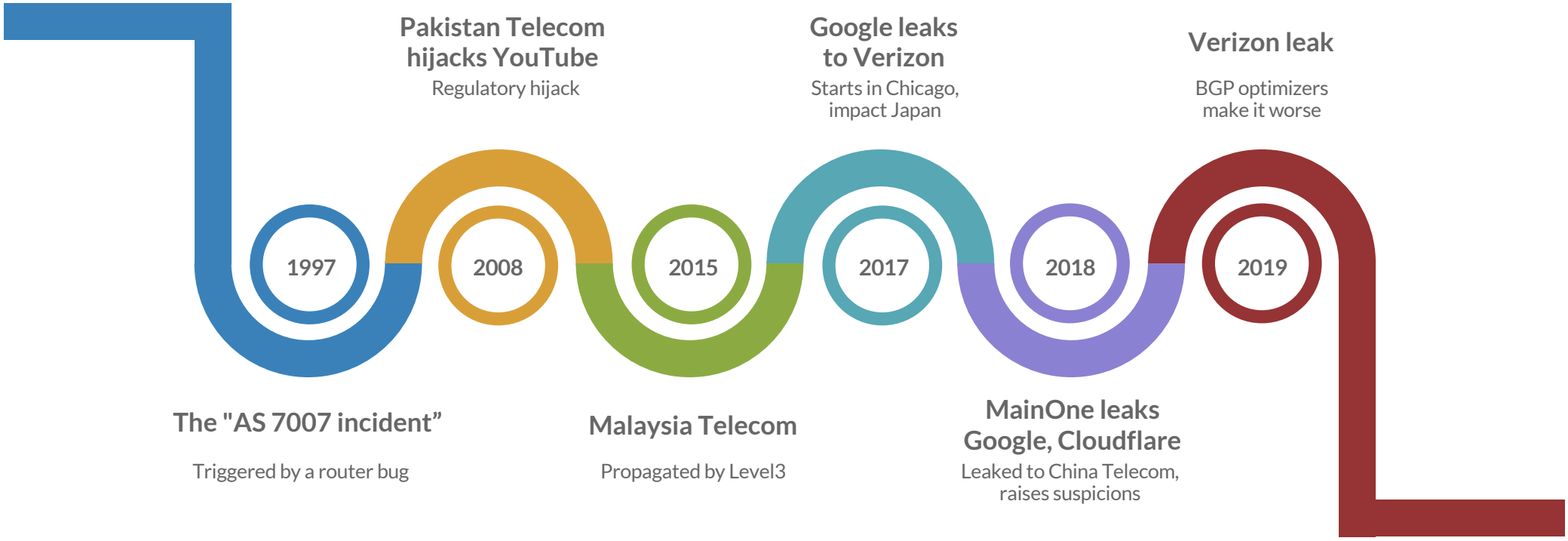
How it looks to the press



CLOUDFLARE

BGP

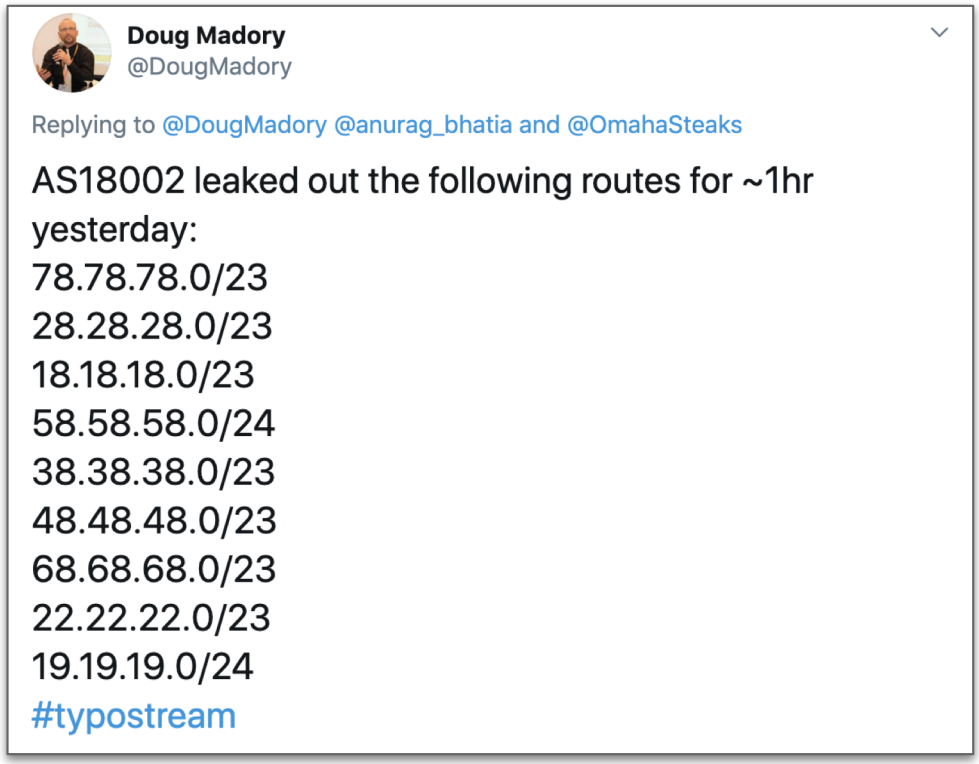
BGP's timeline of leaks



CLOUDFLARE

February 5th, 2020 … the beat goes on!

It would seem that it's not safe to run a "*quad address*"!

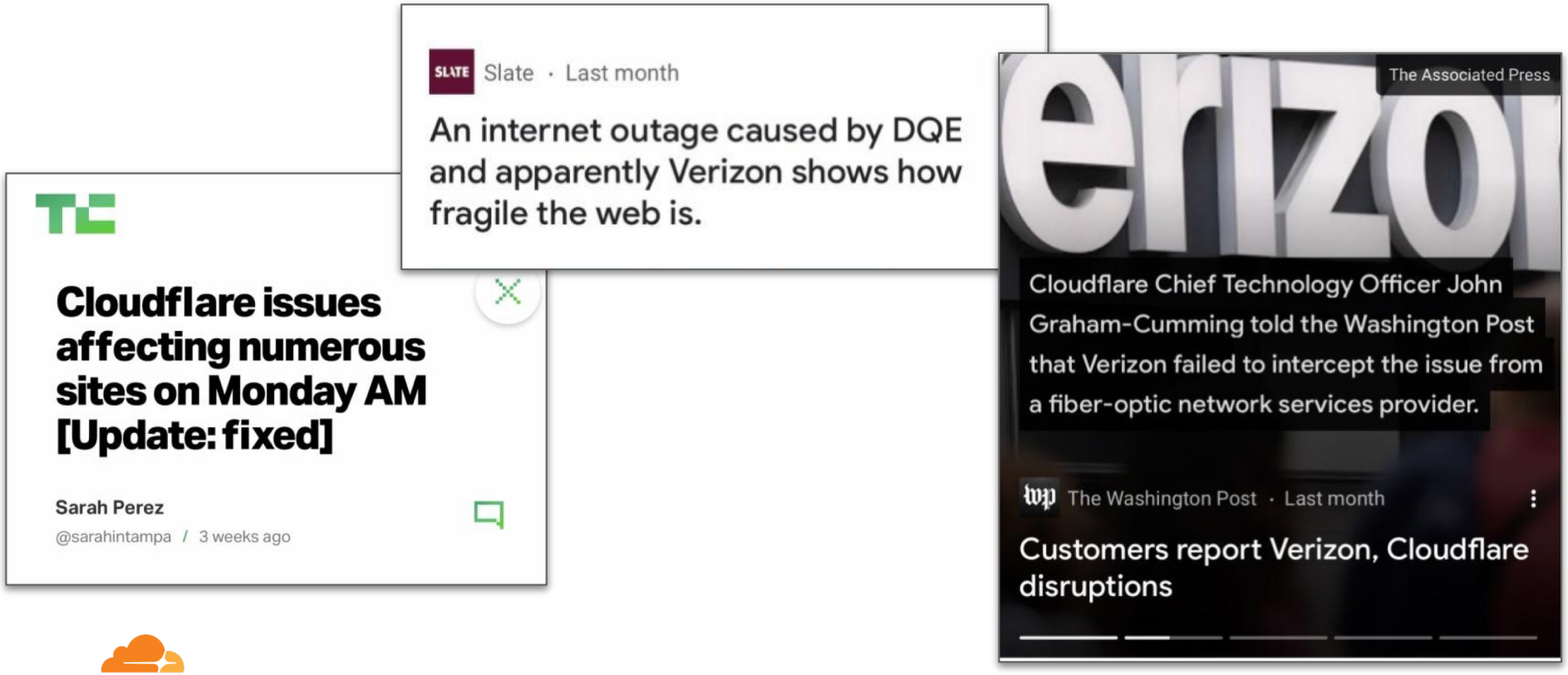


CLOUDFLARE

https://twitter.com/DougMadory/status/1225486918609227777

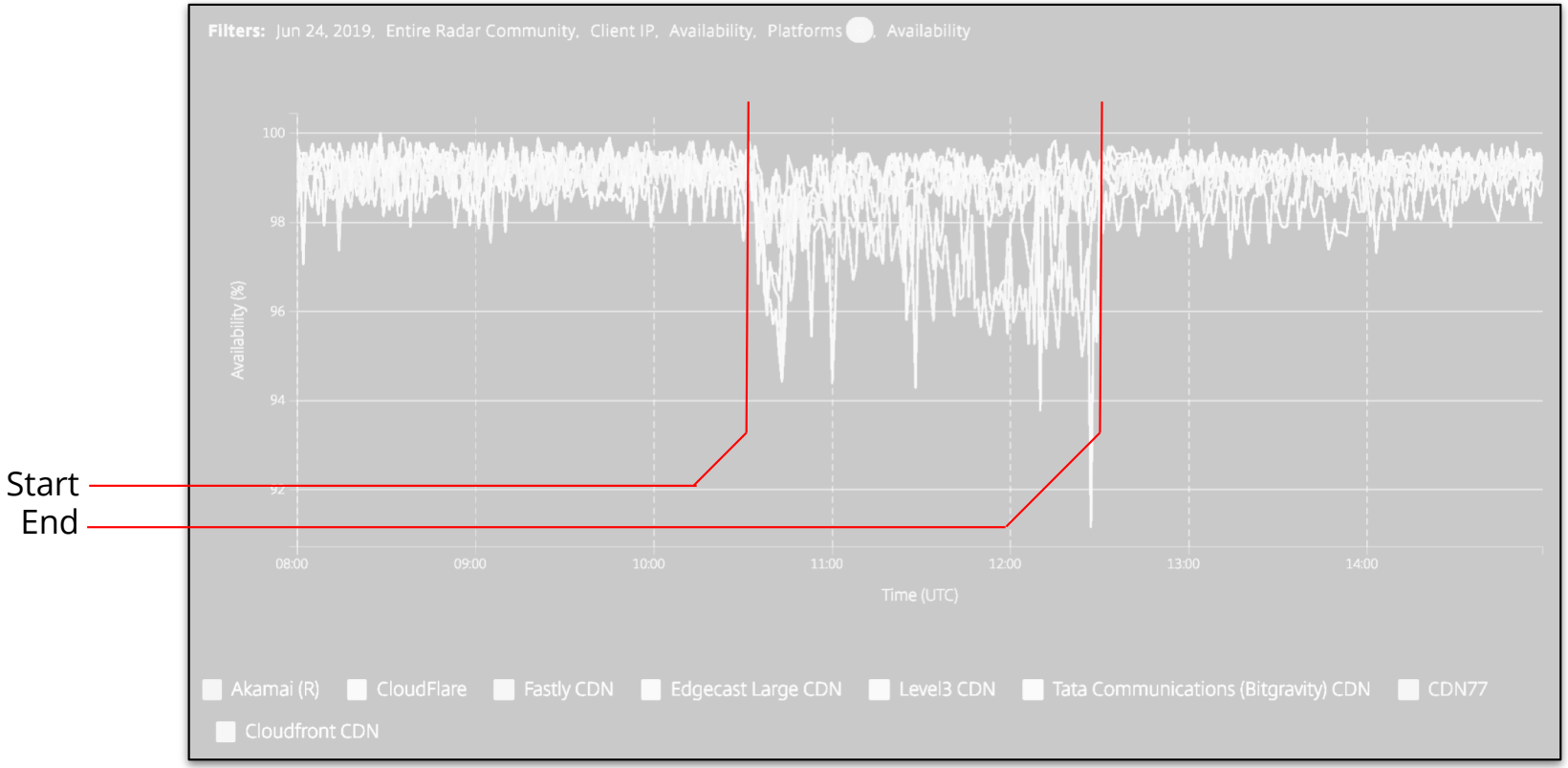
June 24th, 2019, 10:30 UTC

June 24th, 2019 - 10:30 UTC



CLOUDFLARE

June 24th, 2019 leak was widespread

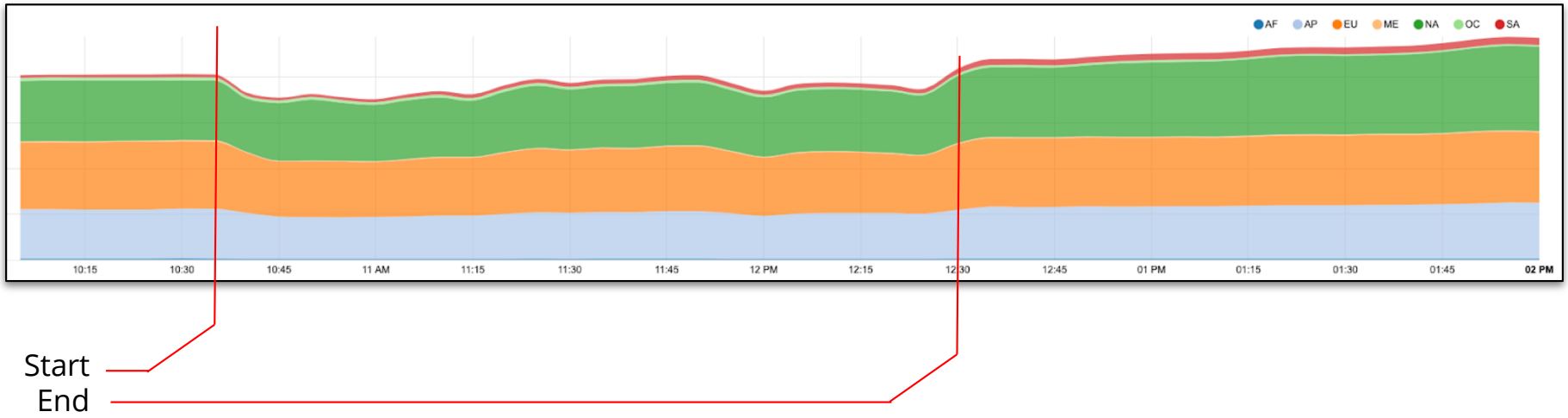


Start

End

Source: Cedexis

June 24th, 2019 Impact on the Cloudflare traffic



CLOUDFLARE

How did it get solved ?

CLOUDFLARE

What is a BGP leak ?

Internet Engineering Task Force (IETF) K. Sriram
Request for Comments: **7908** D. Montgomery
Category: Informational US NIST
ISSN: 2070-1721 D. McPherson
E. Osterweil
Verisign, Inc.
B. Dickson
June 2016

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

CLOUDFLARE

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
```

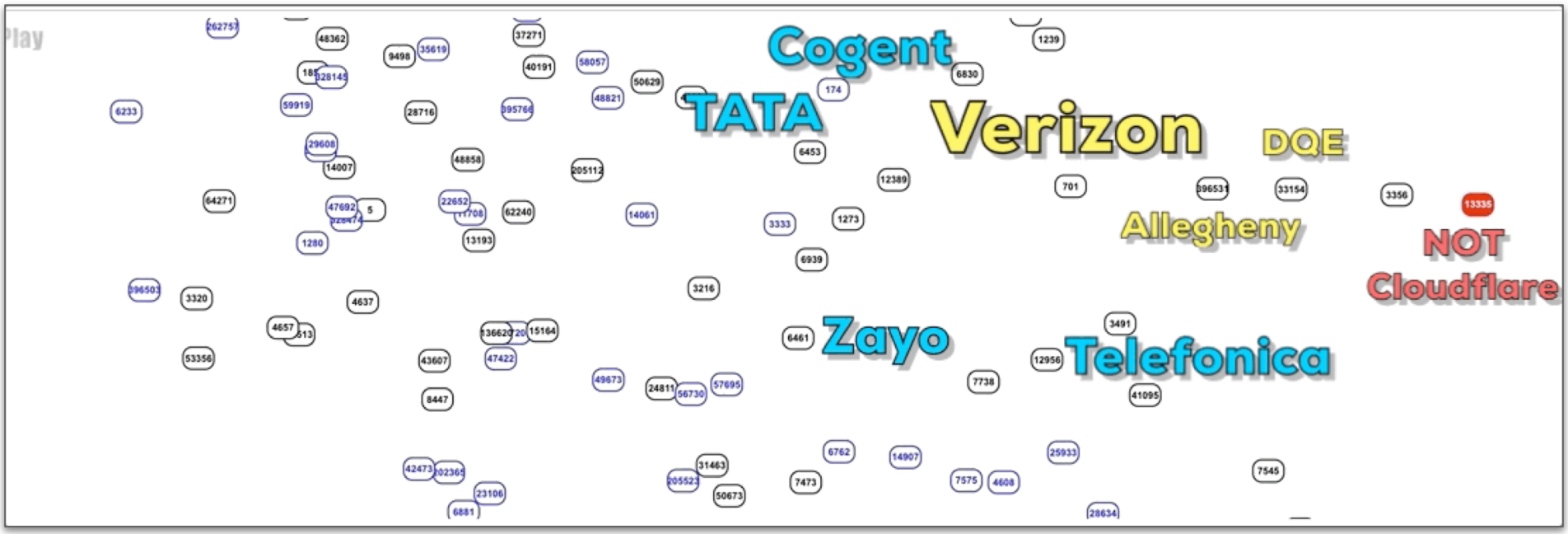
CLOUDFLARE

A very invalid route - step #2

```
Prefix:        104.25.48.0/20
Max Length:    /20
ASN:           13335
Trust Anchor: ARIN
Validity:      Thu, 02 Aug 2018 04:00:00 GMT - Sat, 31 Jul
2027 04:00:00 GMT
Emitted:       Thu, 02 Aug 2018 21:45:37 GMT
Name:          535ad55d-dd30-40f9-8434-c17fc413aa99
Key:           4a75b5de16143adbeaa987d6d91e0519106d086e
Parent Key:    a6e7a6b44019cf4e388766d940677599d0c492dc
Path:          rsync://rpki.arin.net/repository/arin-rpki-
ta/5e4a23ea-...
```

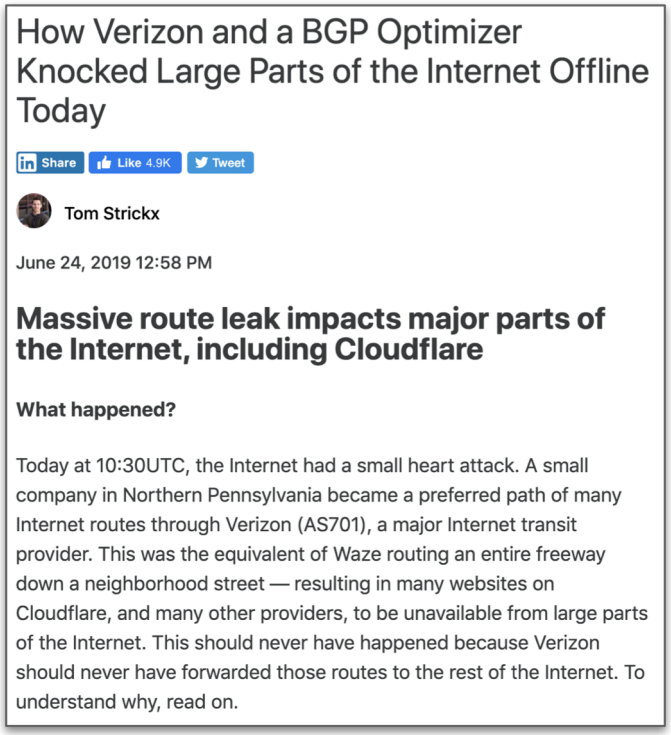
CLOUDFLARE

The disruptive power of Tier 1 providers



CLOUDFLARE

We wrote two blogs about all this



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

Share Like 4.9K Tweet

Tom Strickx

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

Share Like 655 Tweet

Martin J Levy

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 tweeted a quick list of 2,400 ASNs that were affected.

CLOUDFLARE

https://blog.cloudflare.com/how-verizon-and-a-bgp-optimizer-knocked-large-parts-of-the-internet-offline-today/
https://blog.cloudflare.com/the-deep-dive-into-how-verizon-and-a-bgp-optimizer-knocked-large-parts-of-the-internet-offline-monday/

We included all the scripting to show leaks

```
$ # 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
-rw-r--
$
```

```
$ # 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
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 13335-listing-b.txt
$

$ # 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 13335-listing-c.txt
$
```

CLOUDFLARE

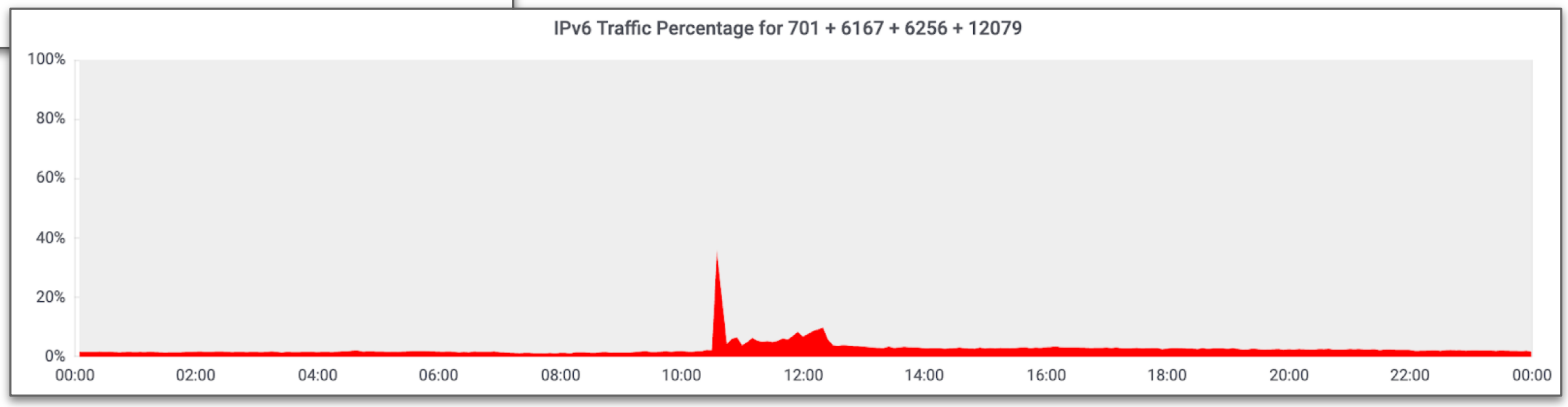
So far this is IPv4 speak - what about IPv6?

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.



CLOUDFLARE

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.

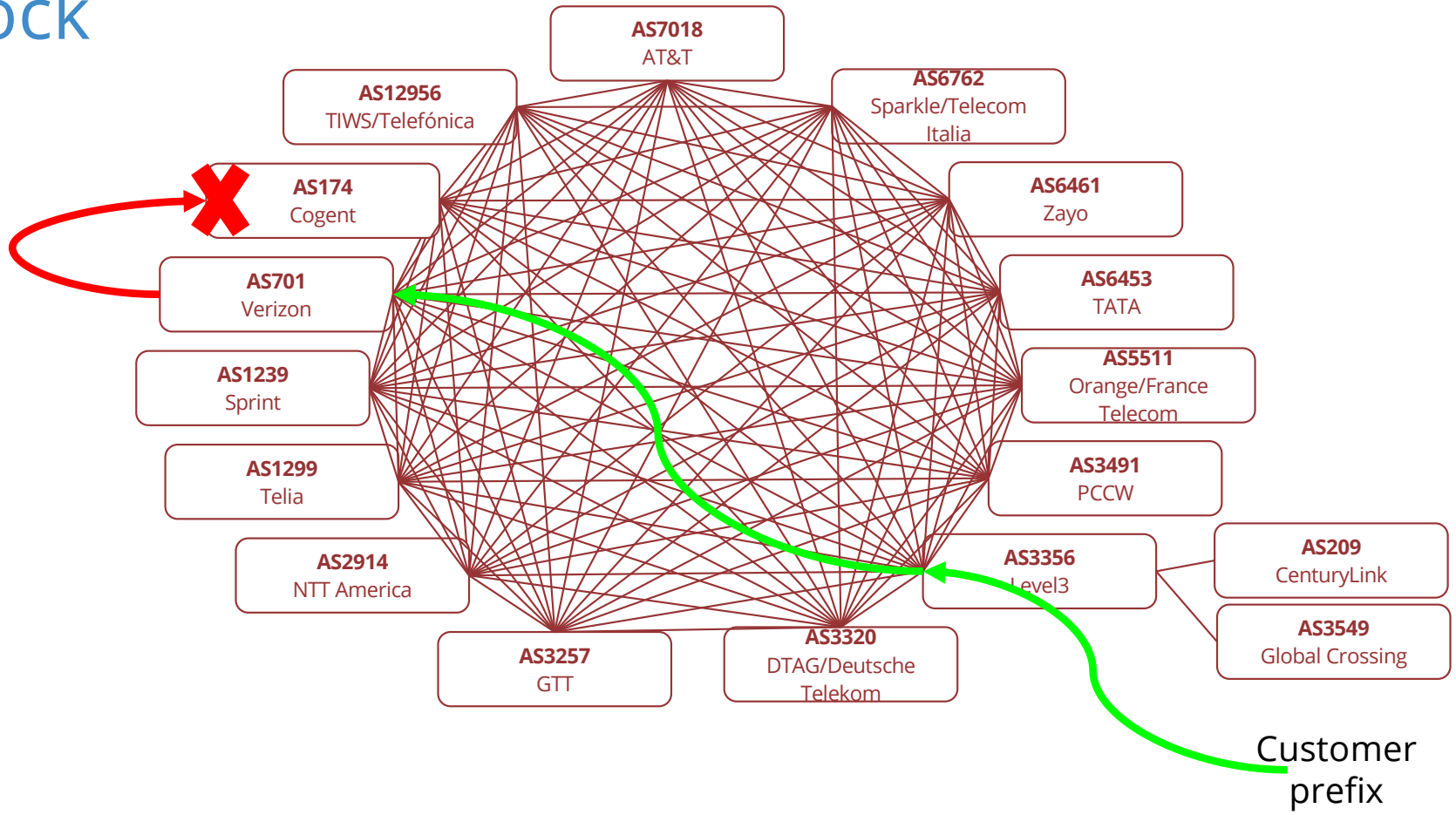
Even if you’re not a Tier1, you can apply this to your customers sessions!

https://archive.nanog.org/sites/default/files/Snijders_Everyday_Practical_Bgp.pdf

CLOUDFLARE

Peerlock

All tier1's have direct interconnection with other tier1's. Financial relationships are not diagramed, this is only routing.



CLOUDFLARE

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)

CLOUDFLARE

BGP route optimizers

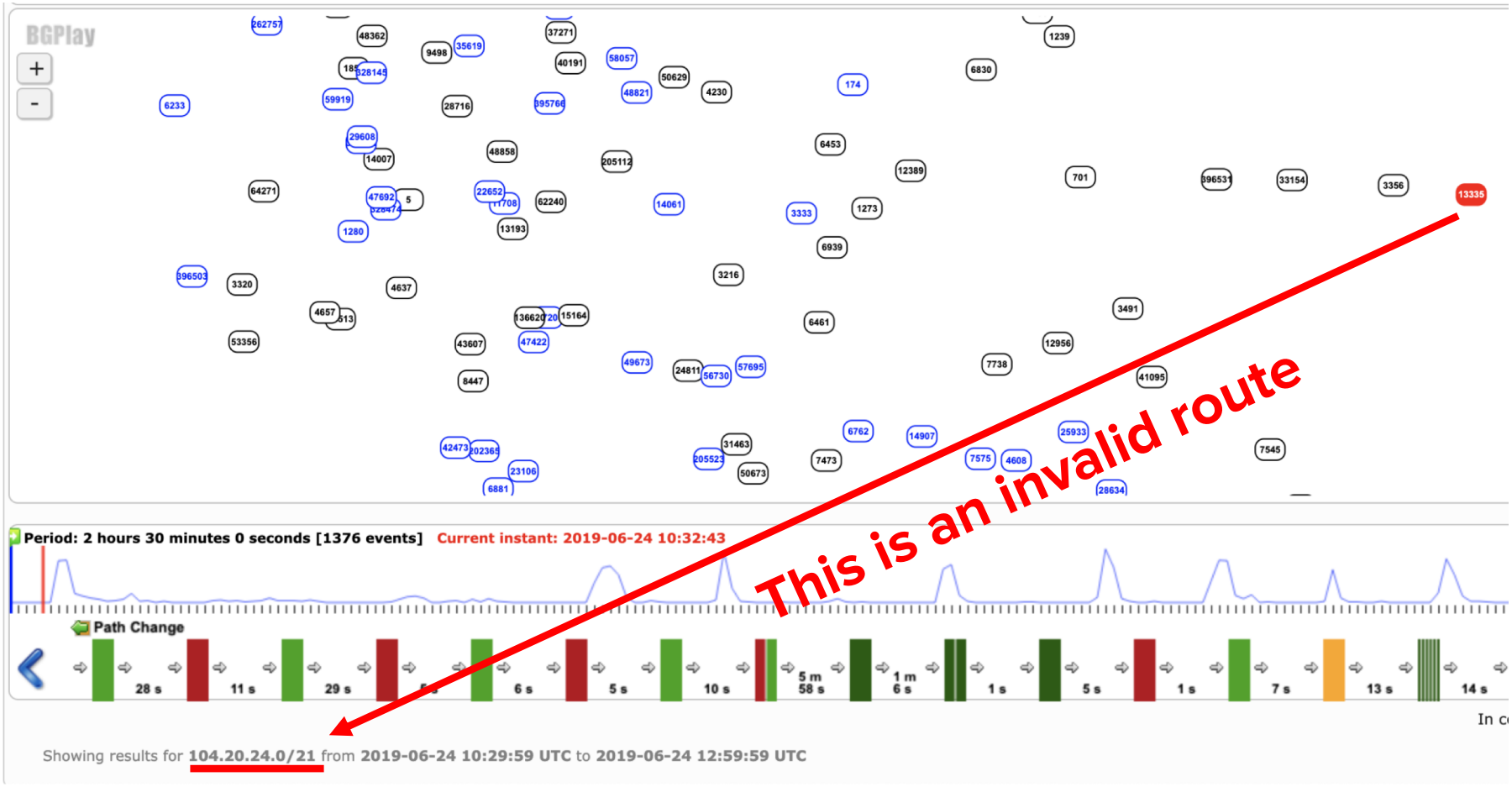
BGP route optimizers make it worse

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.

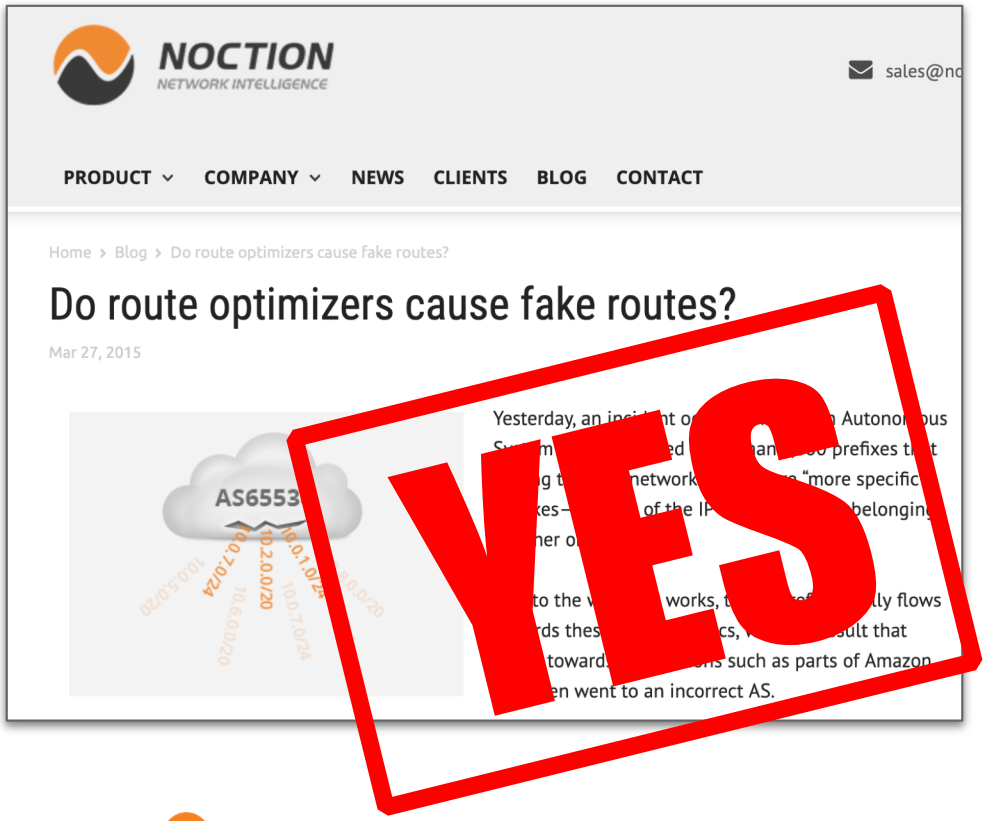
If you leak these “*fake*” routes, you will attract all Internet traffic for these… unless your upstreams filter them.

CLOUDFLARE

BGP optimizers to make it worse



BGP optimizers - our view



CLOUDFLARE

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

CLOUDFLARE

Noction response

Noction responds regarding June/24 route leak.

https://www.noction.com/news/incident-response

> *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/
- 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 ?

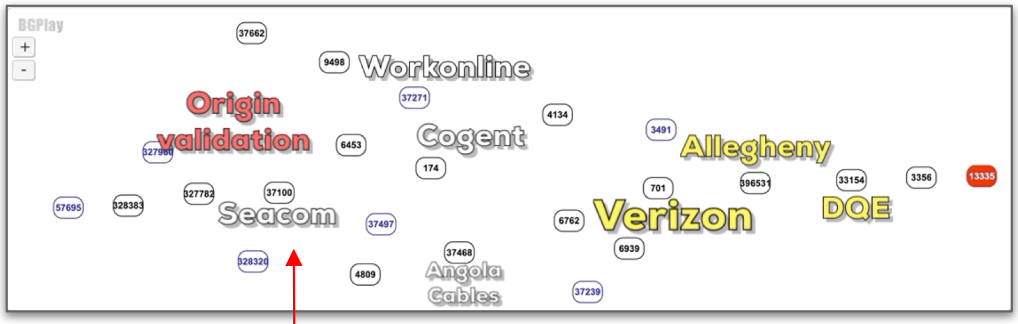
CLOUDFLARE

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

CLOUDFLARE

Route Origin Validation hiccup in Africa



AS37100/Seacom does not use the ARIN TAL; hence routes allocated by ARIN were not protected.

Subject: [JINX.announce] RPKI ROV & Dropping of Invalids - Africa
From: Mark Tinka via jinx-announce <jinx-announce@ispa.org.za>
Date: Tue, Apr 9, 2019 at 5:04 AM

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.

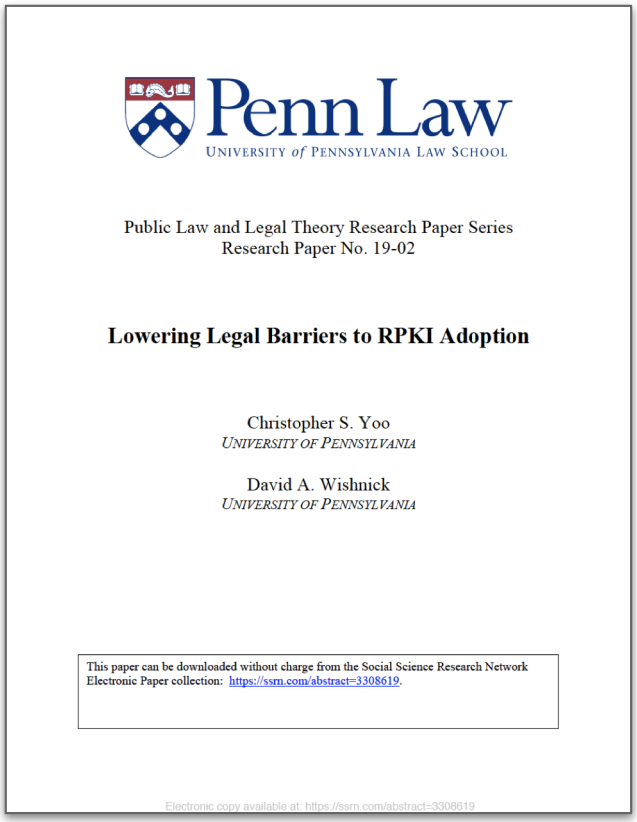
We are still standing by for the 3rd ISP to complete their implementation, and we are certain they will communicate with the community accordingly.

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.

CLOUDFLARE

Lowering Legal Barriers to RPKI Adoption

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3308619



Penn Law
UNIVERSITY of PENNSYLVANIA LAW SCHOOL

Public Law and Legal Theory Research Paper Series
Research Paper No. 19-02

Lowering Legal Barriers to RPKI Adoption

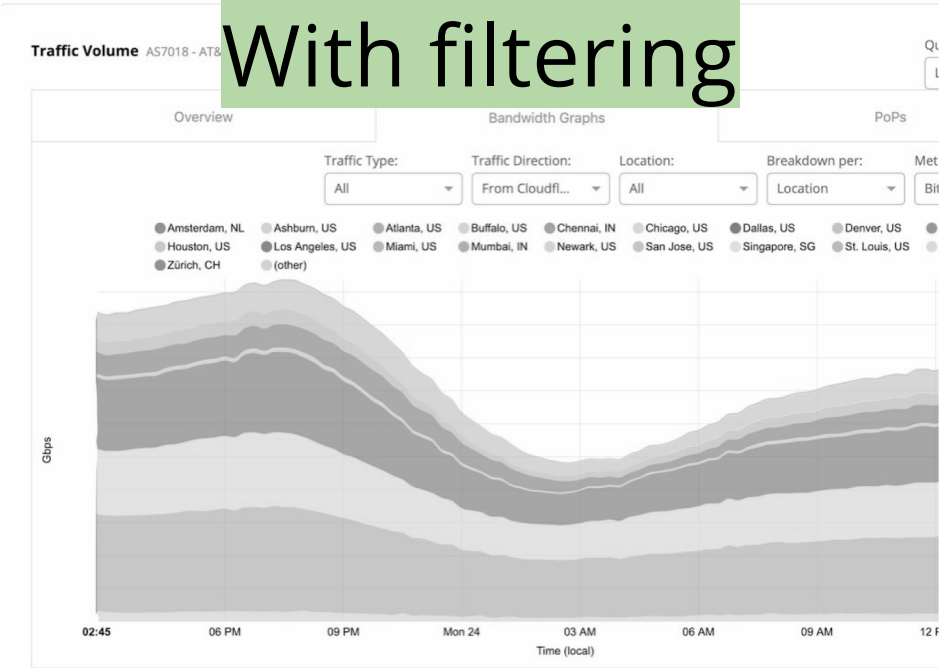
Christopher S. Yoo
UNIVERSITY OF PENNSYLVANIA

David A. Wishnick
UNIVERSITY OF PENNSYLVANIA

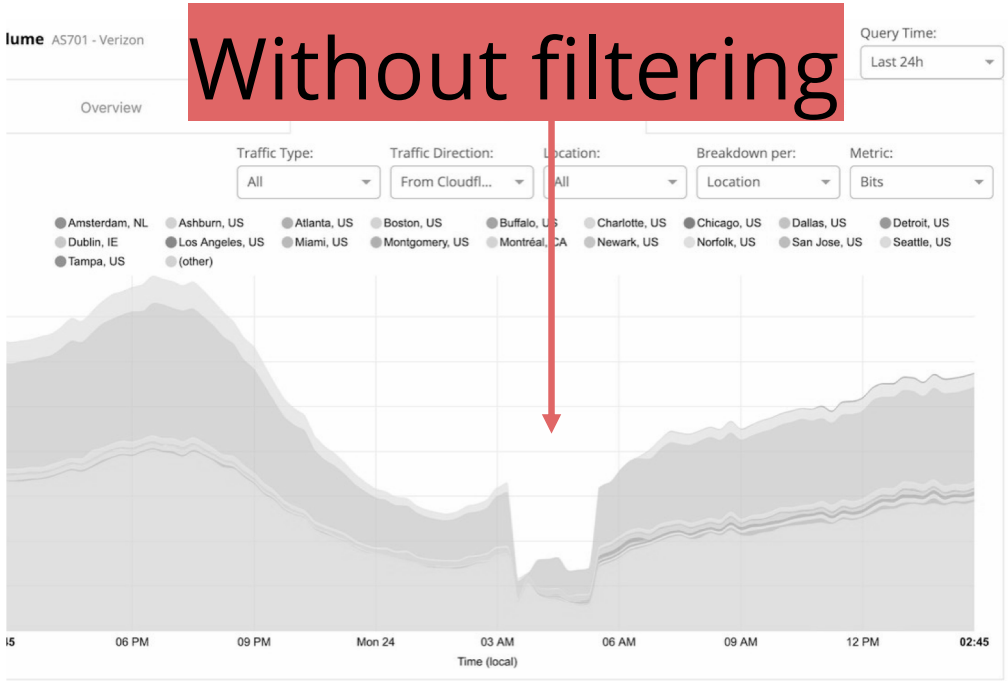
CLOUDFLARE

Deploy RPKI now (Because tomorrow is already too late)

With filtering

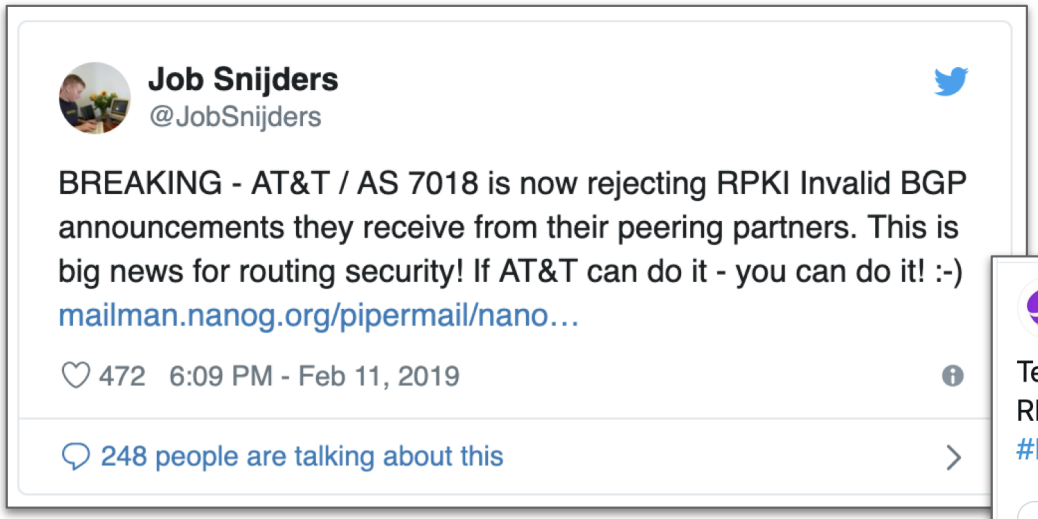


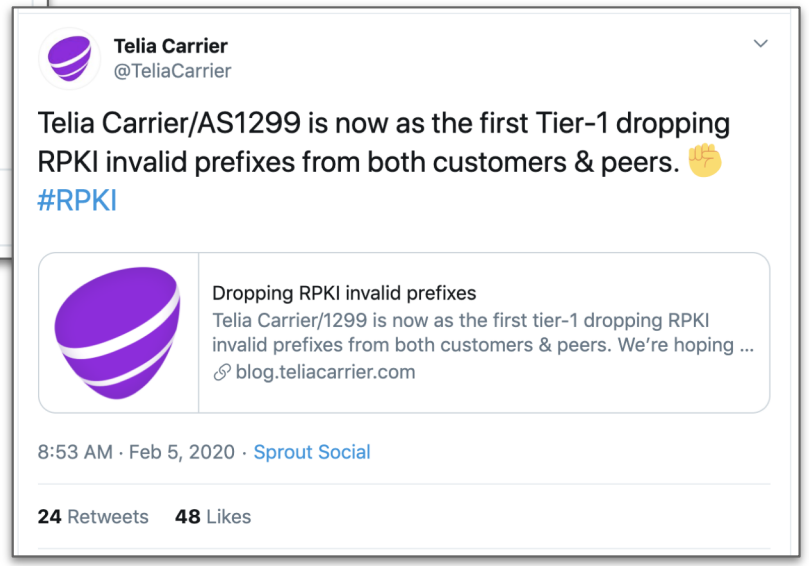
Without filtering



CLOUDFLARE

AS7018/AT&T AS1299/Telia and RPKI





CLOUDFLARE

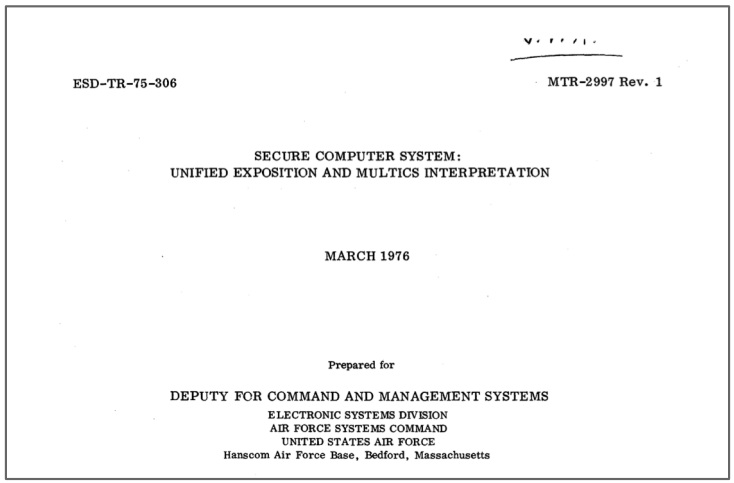
Summary

Questions ?

martin @cloudflare.com
jf @cloudflare.com
tstrickx @cloudflare.com

Additional content

1976 Security in documents



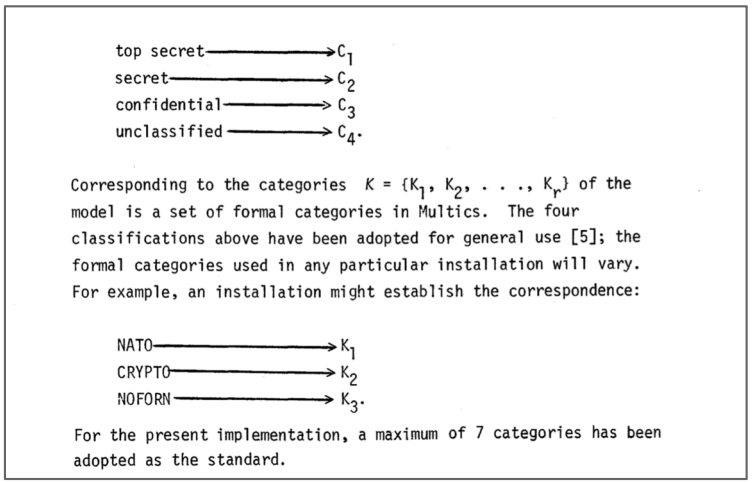
ESD-TR-75-306 MTR-2997 Rev. 1

SECURE COMPUTER SYSTEM:
UNIFIED EXPOSITION AND MULTICS INTERPRETATION

MARCH 1976

Prepared for

DEPUTY FOR COMMAND AND MANAGEMENT SYSTEMS
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
Hanscom Air Force Base, Bedford, Massachusetts



top secret → C_1
secret → C_2
confidential → C_3
unclassified → C_4.

Corresponding to the categories $K = \{K_1, K_2, \ldots, K_r\}$ of the model is a set of formal categories in Multics. The four classifications above have been adopted for general use [5]; the formal categories used in any particular installation will vary. For example, an installation might establish the correspondence:

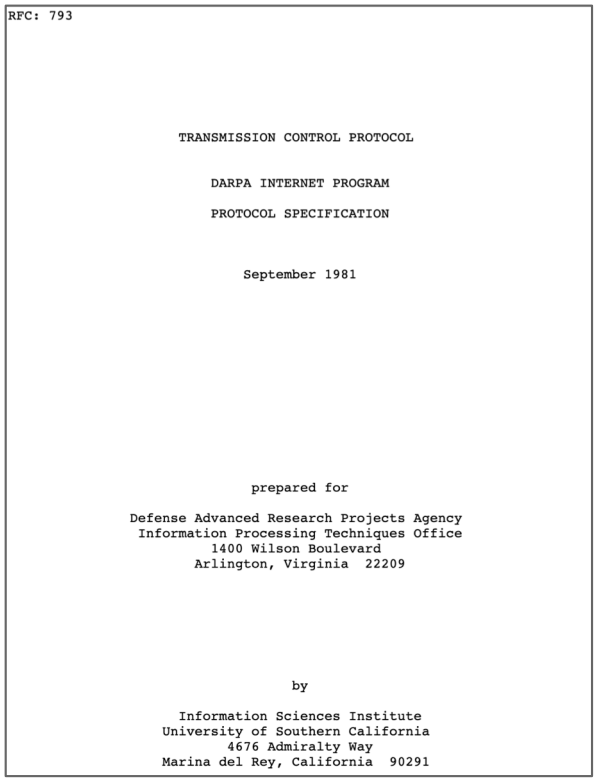
NATO → K_1
CRYPTO → K_2
NOFORN → K_3.

For the present implementation, a maximum of 7 categories has been adopted as the standard.

Security was always being discussed and defined; but mainly in computing

CLOUDFLARE

1981 RFC793 - TCP



RFC: 793

TRANSMISSION CONTROL PROTOCOL

DARPA INTERNET PROGRAM

PROTOCOL SPECIFICATION

September 1981

prepared for

Defense Advanced Research Projects Agency
Information Processing Techniques Office
1400 Wilson Boulevard
Arlington, Virginia 22209

by

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, California 90291

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.

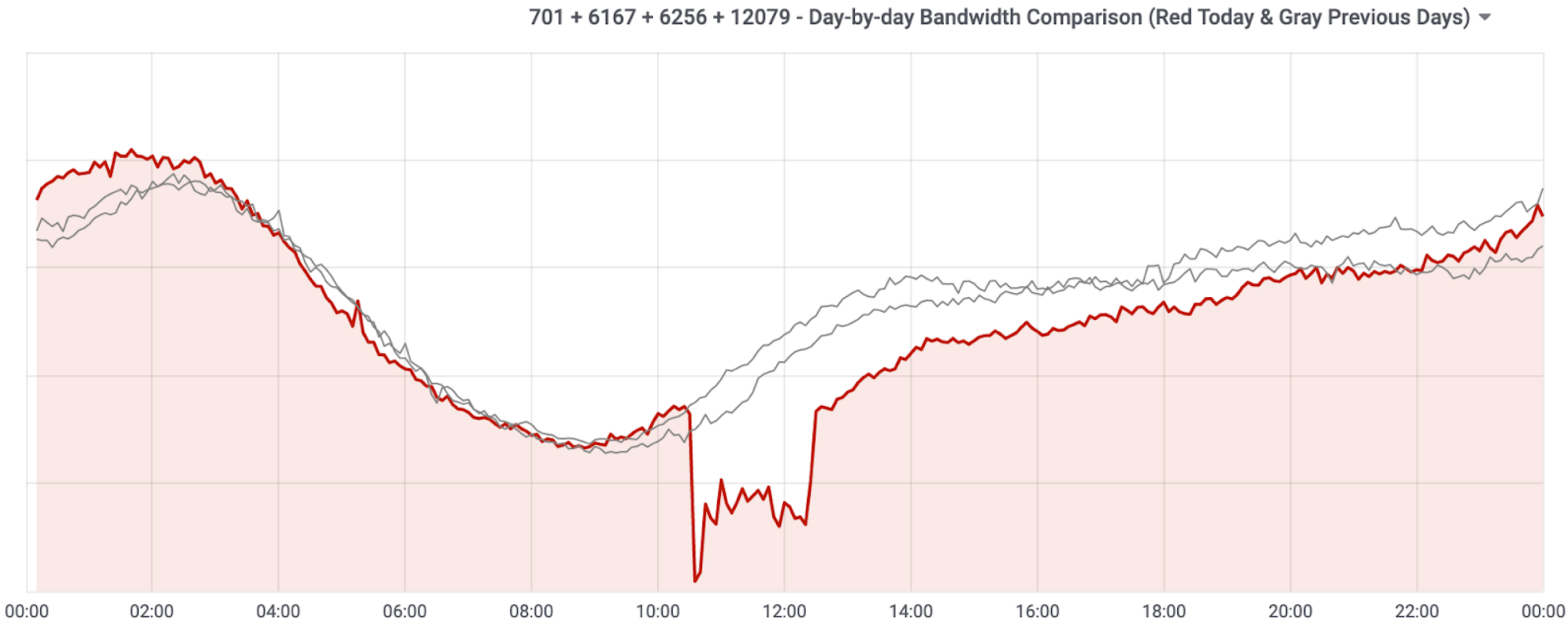
TCP modules which operate in a multilevel secure environment must properly mark outgoing segments with the security, compartment, and precedence. Such TCP modules must also provide to their users or higher level protocols such as Telnet or THP an interface to allow them to specify the desired security level, compartment, and precedence of connections.

RFC793 is the first definition of TCP

2.9. Precedence and Security

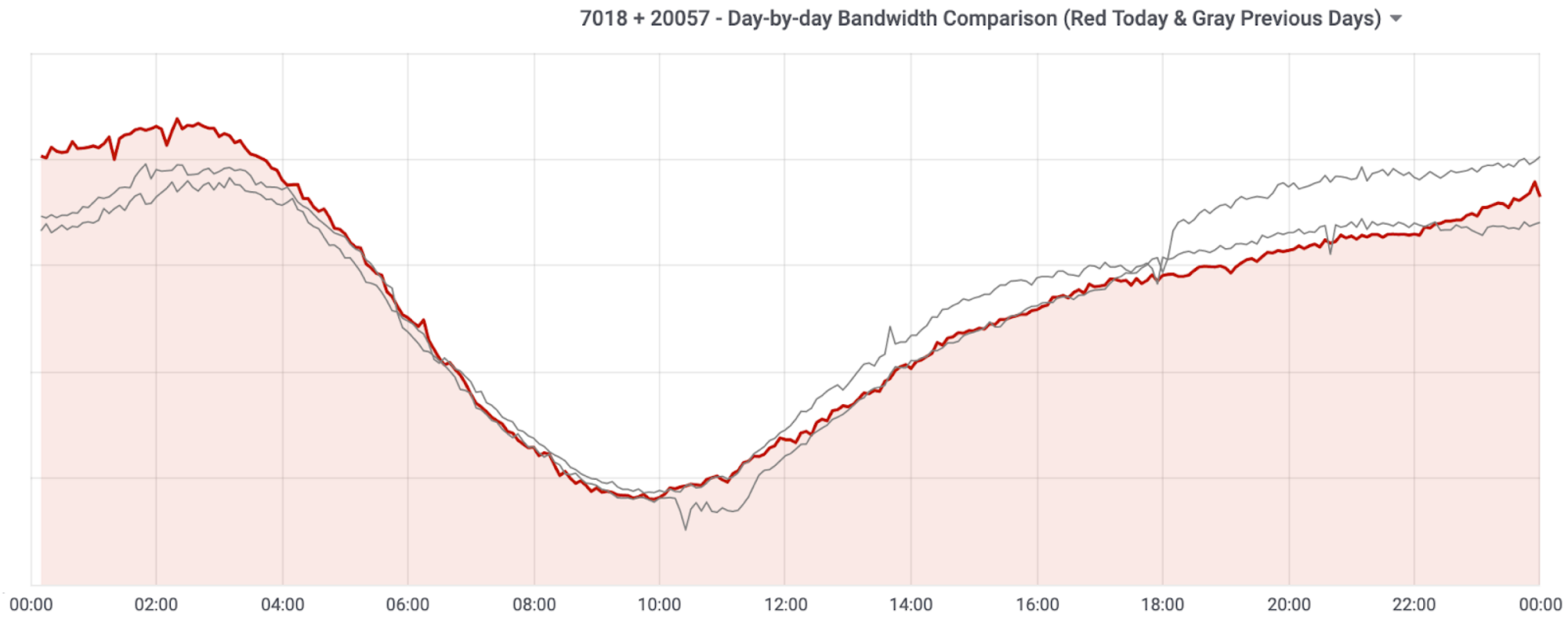
CLOUDFLARE

Without proper filtering



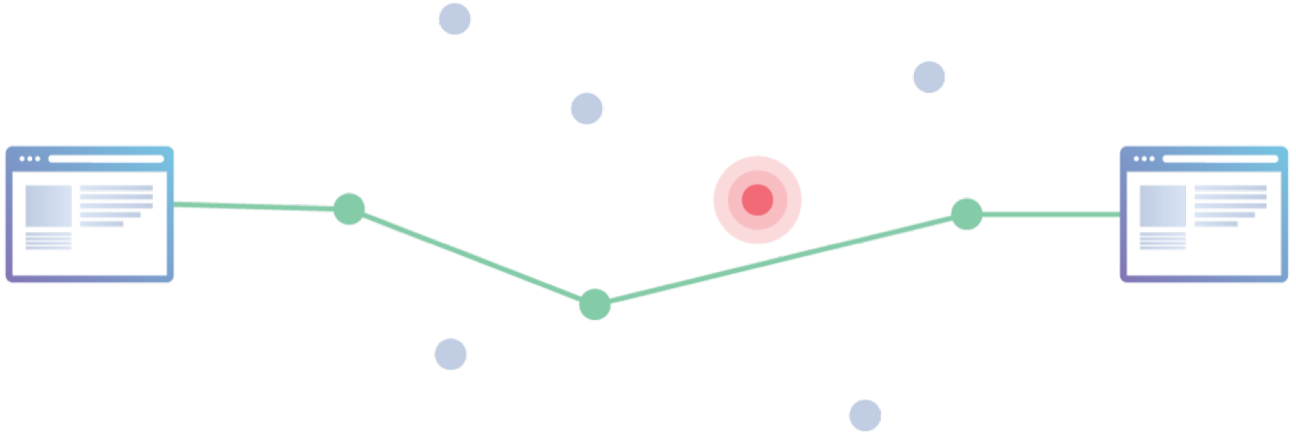
CLOUDFLARE

With proper filtering



CLOUDFLARE

What is BGP?



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

CLOUDFLARE

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

CLOUDFLARE