Alerte en Malaisie, une nouvelle fuite BGP

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 12 juin 2015. Dernière mise à jour le 14 juin 2015

https://www.bortzmeyer.org/bgp-malaisie.html

C'est un accident qui, à cette ampleur, se produit tous les... quoi... trois ou quatre ans sur l'Internet. Un opérateur, en l'occurrence Telekom Malaysia, a laissé fuiter des centaines de milliers de routes Internet, attirant ainsi une quantité de trafic qu'il n'a pas su gérer. Le problème a duré environ deux heures. Du classique, à quelques exceptions près.

Merci à Buck Danny pour la référence :

D'abord, les observations brutes : les premières observations publiques qu'il y avait un problème ont été divers tweets comme celui de Nathalie Rosenberg <https://twitter.com/nrosenberg/status/609282803847532544> à 0855 UTC. Une bonne partie de l'Internet semblait injoignable, et des gens ont commencé à accuser Free ou CloudFlare. En fait, on pouvait voir à plusieurs endroits que la source du problème était dans une crise BGP. Par exemple, les fichiers stockant les annonces BGP à RouteViews <http://www.routeviews.org/> augmentaient brusquement de taille <ftp://archive.routeviews.org/bgpdata/2015.06/UPDATES/>, montrant une forte activité BGP, ce qui n'est pas en général bon signe. De fichiers de moins d'un Mo en temps normal, on est passé à 2,2 Mo à 0845 puis à 14 Mo à 0900 et 19 Mo à 0915. L'examen du contenu de ces fichiers montre la cause du problème : l'AS 4788 (Telekom Malaysia) a laissé fuiter dans les 200 000 routes BGP (au lieu de 1 300 en temps normal, ce qui est déjà beaucoup)... Voici la première annonce erronnée :

```
TIME: 06/12/15 08:43:29
TYPE: BGP4MP/MESSAGE/Update
FROM: 208.51.134.246 AS3549
TO: 128.223.51.102 AS6447
ORIGIN: IGP
ASPATH: 3549 4788 3491 4651 9737 23969
NEXT_HOP: 208.51.134.246
MULTI_EXIT_DISC: 24968
COMMUNITY: 3549:4992 3549:7000 3549:7003 3549:7004 354
9:32344 4788:400 4788:410 4788:415
ANNOUNCE
  1.0.208.0/22
  1.0.212.0/23
  1.1.176.0/22
  ...
```

Aucune de ces préfixes n'est géré par Telekom Malaysia ou ne devrait être annoncé par eux. C'est à tort que leur transitaire Level 3/GlobalCrossing (AS 3549) a accepté ces annonces. (Rappelez-vous qu'un chemin d'AS se lit de droite à gauche, ici 23969 est l'origine.) Normalement, on filtre les annonces de ses clients ou de ses pairs BGP, à partir des IRR. Même si on ne filtre pas sur les données des IRR (souvent mal maintenues), on devrait, au minimum, mettre un nombre maximal de préfixes annoncé et couper la session autrement (si Free l'avait fait, ses clients auraient eu moins de problèmes, voir plus loin). Mais personne ne veut prendre le risque de mécontenter un client. Et puis filtrer demande davantage de travail, alors que la sécurité ne rapporte rien.

Level 3 relayant ce grand nombre de routes, bien des routeurs qui avaient configuré un nombre maximum de préfixes BGP annoncés ont coupé leur session. Ainsi, au RING <https://ring.nlnog.net/>, on voyait <http://sqa.ring.nlnog.net/> :

```
id timestamp raised_by short
2413 2015-06-12 09:33:52 linode01 linode01.ring.nlnog.net: raising ipv4 alarm - 17 new nodes down
2412 2015-06-12 09:32:46 ovh04 ovh04.ring.nlnog.net: raising ipv4 alarm - 43 new nodes down
2411 2015-06-12 09:31:56 ovh03 ovh03.ring.nlnog.net: raising ipv4 alarm - 29 new nodes down
2410 2015-06-12 09:27:49 adix01 adix01.ring.nlnog.net: raising ipv4 alarm - 1 new nodes down
2409 2015-06-12 09:27:46 octopuce01 octopuce01.ring.nlnog.net: raising ipv4 alarm - 20 new nodes down
2408 2015-06-12 09:20:51 berkeley01  berkeley01.ring.nlnog.net: raising ipv4 alarm - 3 new nodes down
```

Notons que tout n'est pas passé par Level 3. Les gens qui "*peeraient*" avec Telekom Malaysia et qui, bien à tort, ne filtraient pas les annonces, ont également reçu les annonces fausses. C'est le cas de Free, ce qui explique l'ampleur de la crise pour les abonnés de Free. Voici un traceroute actuel, montrant le "*peering*" entre Free et Telekom Malaysia :

```
% traceroute www.tm.com.my
traceroute to www.tm.com.my (58.27.84.129), 30 hops max, 60 byte packets
 1  freebox (192.168.2.254)  12.533 ms  12.510 ms  12.498 ms
...
 7  bzn-crs16-1-be1024.intf.routers.proxad.net (212.27.56.149)  25.686 ms  14.741 ms  14.706 ms
 8  th2-9k-1-be1003.intf.routers.proxad.net (78.254.249.97)  12.186 ms  31.208 ms  31.177 ms
 9  yankee-6k-1-po1.intf.routers.proxad.net (212.27.57.14)  115.595 ms  117.136 ms  117.129 ms
10  ash-bo02.tm.net.my (206.126.236.176)  106.776 ms  106.789 ms  115.466 ms
11  10.55.36.116 (10.55.36.116)  366.449 ms  366.450 ms  367.967 ms
12  58.27.84.6 (58.27.84.6)  375.796 ms  375.787 ms  382.950 ms
...
```

Résultat de ces annonces, le trafic filait effectivement en Malaisie comme le montre ce test <https://gist.github.com/anonymous/9ff21a3cbcad452b43ce> :

```
$ mtr -4rwc100 cloudflare.com
Start: Fri Jun 12 11:06:26 2015
HOST: laptop Loss% Snt Last Avg Best Wrst StDev
1.|-- box 0.0% 100 3.4 3.7 3.3 10.5 0.8
2.|-- 129.120.16.109.rev.sfr.net 0.0% 100 28.2 27.7 26.0 39.6 1.4
3.|-- 193.45.66.86.rev.sfr.net 0.0% 100 27.0 27.8 26.0 32.6 0.8
4.|-- 181.45.66.86.rev.sfr.net 0.0% 100 28.6 32.3 26.4 225.8 26.1
5.|-- v3790.poi1-co-1.gaoland.net 0.0% 100 30.6 30.7 27.6 46.4 2.4
6.|-- 54.247.5.109.rev.sfr.net 0.0% 100 38.4 36.2 33.1 42.6 1.5
7.|-- ae1.parigi32.par.seabone.net 2.0% 100 33.8 34.5 32.4 49.1 2.1
8.|-- xe-5-1-2.parigi52.par.seabone.net 0.0% 100 35.1 35.1 33.1 48.0 1.7
9.|-- global-crossing.parigi52.par.seabone.net 72.0% 100 378.4 612.6 378.4 3819. 796.8
10.|-- telekom-malaysia-berhad.xe-0-2-0.ar2.clk1.gblx.net 74.0% 100 399.2 399.3 397.4 412.8 2.8
11.|-- ??? 100.0 100 0.0 0.0 0.0 0.0 0.0
12.|-- 13335.sgw.equinix.com 83.0% 100 460.9 462.5 459.8 480.8 5.1
13.|-- 198.41.214.163 84.0% 100 460.3 460.2 459.0 462.3 0.9
```

Notez que Telekom Malaysia n'a annoncé que 200 000 routes, sur les environ 600 000 qu'on trouve dans l'Internet <http://bgp.he.net/report/netstats> donc environ les deux tiers des sites connectés n'ont pas eu de problème. On voit ici, via DNSmon <https://dnsmon.ripe.net/>, l'effet sur les serveurs DNS de .fr : deux des serveurs perdent environ 1 % des paquets au plus fort de la crise, c'est tout

Les annonces erronnées ont été supprimées vers 1030 UTC mais le trafic BGP n'est complètement revenu à la normale que vers 1200 UTC. (Attention, ce graphique affiche les heures en UTC+2, alors qu'une supervision de l'Internet devrait toujours utiliser UTC.)

Une des originalités de cette fuite est que l'origine du chemin d'AS n'était pas le fuiteur, Telekom Malaysia. La plupart du temps, le fuiteur émet les routes comme s'il en était l'origine et des techniques comme RPKI+ROA <https://www.bortzmeyer.org/securite-routage-bgp-rpki-roa.html> détectent l'usurpation. Ici, au contraire, l'origine... originelle a été préservée et RPKI+ROA n'auraient donc rien vu. On voit ici une annonce Telekom Malaysia sur un routeur qui valide et on note qu'il a validé l'annonce :

```
193.0.0.0/21
                    [BGP/170] 00:20:29, MED 1000, localpref 150
                    AS path: 3549 4788 12859 3333 I, validation-state: valid
                  > to 64.210.69.85 via xe-1/1/0.0
```

Alors, BGP mérite-t-il le surnom de « Bordel Gateway Protocol » que Nathalie Rosenberg <https://twitter.com/nrosenberg/status/609301968830496768> lui avait donné ? Il est vrai que l'acceptation sans discussions d'annonces par les autres opérateurs est une source de fragilité. Mais les ingénieurs réagissent rapidement <https://www.bortzmeyer.org/securite-bgp-et-reaction-rapide.html> et s'adaptent et la panne, finalement, n'aura pas duré longtemps.

D'autres articles sur ce sujet :

— Un article technique de BGPmon <http://www.bgpmon.net/massive-route-leak-cause-internet-slo
>.
— Une autre, par Dyn (ex-Renesys) <http://research.dyn.com/2015/06/global-collateral-damage-of-
>, qui étudie bien le cas de Free, et le rôle qu'a joué leur "*peering*".
— S'il s'est passé quelque chose sur l'Internet, il y a un article de Geoff Huston. C'est le cas ici <https://labs.ripe.net/Members/gih/more-leaky-routes>.
— Une critique de Level 3 sur la liste NANOG <http://mailman.nanog.org/pipermail/nanog/2015-June/076225.html>, suivie d'une bonne discussion.
— Un bon résumé en français dans Next Inpact <http://www.nextinpact.com/news/95399-internet-parti
htm> et un autre dans le Monde <http://www.lemonde.fr/pixels/article/2015/06/12/pourquoi-internet-etait-un-peu-lent-vendredi-matin_4653156_4408996.html>.
— Le communiqué officiel de Telekom Malaysia <https://www.tm.com.my/OnlineHelp/Announcement/Pages/INTERNET-SERVICES-DISRUPTION-12-June-2015.aspx> (l'horreur "*corporate*" à son sommet). Celui de Level 3 <https://twitter.com/Level3/status/609353696787496960> n'est pas mal non plus.