BGP et les trous noirs

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 17 juin 2011

https://www.bortzmeyer.org/bgp-trou-noir.html

Le 16 juin, le routeur de Renater au point d'échange Sfinx est devenu un **trou noir**, un routeur qui annonce des routes mais ne transmet pas les paquets ensuite. Pourquoi les mécanismes de secours normaux de BGP n'ont-ils servi à rien ?

Le protocole BGP (RFC 4271 [1]) permet les échanges de routes entre opérateurs Internet. Chaque routeur annonce à ses pairs les routes qu'il connaît directement ou indirectement (par exemple « Je sais joindre 2001:db8:42::/48 »). Si les pairs sélectionnent ces routes, ils enverront ensuite les paquets à destination de ces préfixes au routeur qui les a annoncés. Ce système est normalement très robuste. Si un routeur stoppe, plante, est débranché, ou est physiquement détruit, la session BGP avec les pairs stoppe et ceux-ci refont tourner l'algorithme de sélection de routes et choisissent d'autres routes. C'est en bonne partie sur ce mécanisme que repose la résistance de l'Internet aux pannes.

Mais ce mécanisme n'a pas fonctionné le 16 juin (ticket Renater n° 2214325). Pour une raison inconnue, le routeur fonctionnait toujours, maintenait les sessions BGP mais ne transmettait plus les paquets, qui étaient simplement jetés. On parle de « trou noir » bien qu'on pourrait aussi dire que le routeur est un « allumeur » : il attire le trafic mais n'assume pas son rôle par la suite. Symptôme avec traceroute : juste avant ou juste après le routeur en question (selon la façon dont il traite les paquets dont le TTL a expiré), on ne voit plus que des étoiles :

```
traceroute vers rigolo.nic.fr (2001:660:3003:2::4:20) de 2001:..., port 33434, du port 45934, 30 sauts max, 60 o
...
3  te2-2-72-nb-stdenis-2.ipv6.nerim.net (2001:7a8:1:72::2)  0.487 ms  0.450 ms  0.465 ms
4  te2-2-94-nb-voltaire-1.ipv6.nerim.net (2001:7a8:1:94::1)  0.658 ms  0.701 ms  0.699 ms
5  te2-4-20-nb-voltaire-2.ipv6.nerim.net (2001:7a8:1:20::2)  0.712 ms  0.726 ms  0.695 ms
6  renater-th2.sfinx.tm.fr (2001:7f8:4e:2::103)  2.916 ms  0.900 ms  1.333 ms
7  * * *
8  * * *
9  * * *
```

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc4271.txt

(Merci à Laurent Dolosor pour le traceroute.)

BGP ne fournit pas de solution à ce problème. Tant que la session fonctionne, il considère que le routeur tiendra ses promesses et transmettra les paquets. Le fait d'être "*multi-homé*" n'aide donc pas, puisque les routes alternatives ne seront pas essayées. Les seuls mécanismes sont, chez les pairs du routeur défaillant, de détecter l'absence de trafic et/ou la non-transmission des paquets et de fermer les sessions BGP. Ce n'est pas facile à automatiser (d'autant plus qu'on risque de couper des sessions BGP à tort si le système de détection est trop sensible). Cela ne suffit pas si un serveur de routes est utilisé, il faut alors aussi couper la session avec le serveur de routes (ce qui est très violent car cela fait perdre plein d'autres routes que celles du routeur allumeur), soit filtrer spécifiquement les annonces de l'opérateur impacté (par exemple avec `route-map` sur IOS ou `route-filter` sur JunOS, mais attention, cette méthode impose de penser à supprimer les filtres une fois le problème disparu).

L'idéal est que l'opérateur du routeur à problèmes éteigne la machine coupable, ramenant ainsi le problème à un cas connu (pair BGP arrêté).

Un article sur un problème similaire (peut-être inspiré par la même panne) est (original en anglais) « "*When Null0 and BGP May Cause Problems*" <http://gandikitchen.net/post/2011/06/20/When-Null0-and-BGP-May-Cause-Problems> » et « Quand Null0 et BGP peuvent causer problème <http://lacuisinedegandi.net/post/2011/06/20/Quand-Null0-et-BGP-peuvent-causer français).