Annonces BGP plus spécifiques : bien ou mal ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 4 septembre 2017

https://www.bortzmeyer.org/bgp-more-specifics-ok.html

L'article de Geoff Huston « *"BGP More Specifics : Routing Vandalism or Useful ?"* <https://blog.apnic.net/2017/06/26/bgp-specifics-routing-vandalism-useful/> » aborde une question qui a toujours déclenché de chaudes discussions dans les forums d'opérateurs réseaux : est-ce que les opérateurs qui, au lieu d'annoncer un et un seul préfixe IP très général, annoncent plusieurs sous-préfixes, sont des salauds ou pas ? Est-ce l'équivalent de polluer et de taguer ? Après une étude soignée et pas mal de graphiques, il apporte une réponse originale.

Comme d'habitude, je vais commencer par un petit rappel. Le protocole BGP, normalisé dans le RFC 4271 [1], est le protocole standard de distribution des routes dans l'Internet. Une route est **annoncée** sous forme d'un préfixe IP, par exemple `2001:db8::/32`, et les routeurs avec qui on communique relaieront cette annnonce, jusqu'à ce que tout l'Internet soit au courant. Le préfixe annoncé inclut une longueur (32 bits dans l'exemple ci-dessus). Comme des expressions comme « grand préfixe » sont ambiguës (parle-t-on de la longueur du préfixe, ou bien du nombre d'adresses IP qu'il contient ?), on parle de préfixes **plus spécifiques** (longueur plus grande, moins d'adresses IP) ou **moins spécifiques** (longueur plus réduite, davantage d'adresses IP). Ainsi, `2001:db8:42:abcd::/64` est plus spécifique que `2001:db8:42::/48`.

Si un opérateur a le préfixe `2001:db8::/32`, et qu'il annonce en BGP non seulement ce préfixe mais également `2001:db8:42::/48` et `2001:db8:cafe::/48`, on dit qu'il annonce des préfixes plus spécifiques. Comme toute route va devoir être stockée dans **tous** les routeurs de la DFZ, cet opérateur qui annonce **trois** routes au lieu d'une seule est vu par certains comme un gougnafier qui abuse du système (et, là, on cite en général la fameuse « *"Tragedy of the Commons"* <https://www.bortzmeyer.org/communs-et-nobel.html> »). Cet opérateur force tous les routeurs de la DFZ à consacrer des ressources (CPU, mémoire) à ses trois préfixes, et cela gratuitement, alors qu'il pourrait se contenter d'un seul préfixe, qui englobe les trois en question.

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

Le problème est d'autant plus complexe que l'Internet est un réseau décentralisé, sans chef et sans police, et que personne ne peut faire respecter des règles mondiales. Donc, même si on pense que les plus spécifiques sont une mauvaise chose, comment les interdire en pratique ?

En général, les opérateurs estiment que les annonces plus spécifiques sont une mauvaise chose. Néanmoins, pas mal de gens en font. Huston regarde de plus près pourquoi, et quelles sont les conséquences. Les plus spécifiques peuvent être utiles pour leur émetteur, car la transmission de paquets IP est fondée sur le principe du « **préfixe le plus spécifique** » : si un routeur IP a deux routes pour la destination `2001:db8:cafe::1:443`, la première pour le préfixe `2001:db8::/32`, et la seconde pour le préfixe `2001:db8:cafe::/48`, il utilisera toujours la seconde, la plus spécifique. (Si, par contre, il doit envoyer un paquet à `2001:db8:1337::bad:dcaf`, il utilisera la première route, puisque la seconde ne couvre pas cette destination.) L'utilisation de préfixes plus spécifiques peut donc être un moyen de déterminer plus finement le trajet qu'emprunteront les paquets qui viennent vers vous.

Quel est le pourcentage de ces « plus spécifiques », d'ailleurs ? Huston mesure environ 50 % des routes IPv4 et 40 % en IPv6. C'est donc une proportion non négligeable. La question est « peut-on / doit-on en réduire le nombre ? Est-ce possible ? Quels sont les coûts et les bénéfices ? »

Pour répondre à cette question, Huston développe une taxonomie des annonces plus spécifiques, avec des exemples réels. (Les exemples ci-dessous ont été vérifiés mais attention, le "*looking glass*" typique n'indique que le préfixe le plus spécifique ; si on n'a pas soi-même une copie de la DFZ et les outils pour l'analyser, il faut utiliser un service qui affiche également les préfixes englobants - je me sers de RIPE Stat <`https://stat.ripe.net/`> **et** un "*looking glass*" - j'utilise celui de Hurricane Electric <`https://lg.he.net/`>.) Huston distingue trois cas. Le premier est celui de « percement d'un trou » : l'annonce plus générale n'envoie pas au bon AS et on fait donc une annonce plus spécifique avec un AS d'origine différent. On voit ainsi, pour `72.249.184.0/24`, sur RIPE Stat :

```
Originated by: AS394094 (valid route objects in LEVEL3 and RADB)

Covering less-specific prefixes:
72.249.128.0/18 (announced by AS30496)
72.249.184.0/21 (announced by AS36024)

Showing results for 72.249.184.0/24 as of 2017-09-03 08:00:00 UTC
```

Le /21 est annoncé par l'AS 36024 alors qu'il y a une annonce plus spécifique pour le /24, avec un autre AS, 394094. Huston note que cela pourrait être fait en annonçant un préfixe par /24 mais cela créerait davantage d'entrées dans la table de routage que ce percement de trou avec un plus spécifique.

Le second cas d'annonce plus spécifique est l'ingénierie de trafic. L'AS d'origine est le même, mais l'annonce est différente sur un autre point, par exemple par le "*prepending*" (la répétition du même AS dans l'annonce, pour rendre le chemin plus long et donc décourager l'usage de cette route), ou bien en envoyant les deux préfixes, le plus spécifique et le moins spécifique, à des opérateurs différents :

```
Originated by: AS4775 (valid route objects in RADB and NTTCOM)

Covering less-specific prefix: 1.37.0.0/16 (announced by AS4775)

Showing results for 1.37.27.0/24 as of 2017-09-03 08:00:00 UTC
```

Cette fois, l'AS d'origine est le même. Mais on note sur le "*looking glass*" que les annonces sont différentes :

```
Prefix: 1.37.0.0/16
          AS_PATH: 4766 4775

Prefix: 1.37.27.0/24
         AS_PATH: 4775
```

Le préfixe plus spécifique n'est pas annoncé aux mêmes opérateurs.

Enfin, le troisième et dernier cas est celui du recouvrement complet : l'annonce plus spécifique est apparemment inutile puisque il existe une annonce moins spécifique qui a exactement les mêmes caractéristiques. Pourquoi ferait-on cela ? Il peut s'agir de négligence ou d'incompétence mais il existe aussi une raison rationnelle : limiter les conséquences d'un détournement BGP.

Dans un tel détournement, l'attaquant annonce un préfixe qui n'est pas le sien, pour perturber l'accès à la victime, ou pour détourner son trafic vers l'attaquant (les cas les plus connus sont accidentels, comme le fameux détournement de YouTube par le Pakistan <https://www.bortzmeyer.org/pakistan-pirate-youtube.html>, mais cela peut aussi se faire délibérement). Si l'annonce normale est un /22, et que l'attaquant annonce aussi un /22, une partie de l'Internet verra toujours l'annonce légitime (les routeurs BGP ne propagent, pour un préfixe donné, que la meilleure route) et ne sera donc pas affectée. Pour qu'une attaque réussisse bien, dans ces conditions, il faut que l'attaquant soit très bien connecté, avec de nombreux partenaires BGP. Par contre, avec cette annonce légitime en /22, un attaquant qui enverrait une annonce pour un /24 verrait son annonce propagée partout (puisqu'il s'agirait d'un préfixe différent). Et, au moment de la transmission des paquets, les routeurs utiliseront la route la plus spécifique, donc le /24. Ainsi, un attaquant mal connecté pourra toujours voir son annonce acceptée et utilisée dans tout l'Internet.

Le fait d'annoncer un recouvrement complet (quatre /24 en plus du /22) protège donc partiellement contre cette technique. Huston rejette cet argument en disant qu'il ne s'agit que d'une protection partielle, mais je ne suis pas d'accord : une protection partielle vaut mieux que pas de protection du tout et, en sécurité, il est courant que les solutions soient partielles.

Un exemple est 1.0.4.0/22. On voit les quatre préfixes plus spécifiques, avec exactement le meme contenu :

```
Prefix: 1.0.4.0/22
          AS_PATH: 4826 38803 56203

Prefix: 1.0.4.0/24
         AS_PATH: 4826 38803 56203
Prefix: 1.0.5.0/24
         AS_PATH: 4826 38803 56203
Prefix: 1.0.6.0/24
         AS_PATH: 4826 38803 56203
Prefix: 1.0.7.0/24
         AS_PATH: 4826 38803 56203
```

Les annonces plus spécifiques forment donc la moitié (en IPv4) des annonces dans la DFZ. Mais ce pourcentage augmente-t-il ou diminue-t-il ? Est-ce que la situation s'aggrave ? Huston utilise les données BGP accumulées depuis dix ans (si vous voulez faire pareil, les annonces BGP sont archivées et disponibles, entre autres à RouteViews <http://archive.routeviews.org/>). Eh bien, il n'y a pas de changement en IPv4 : le pourcentage de 50 % des annonces étant des annonces plus spécifiques qu'une

annonce existante n'a pas changé depuis dix ans (mais la proportion des trois cas a changé : lisez l'article). En revanche, il est passé de 20 à 40 % en IPv6 dans le même temps, mais je ne suis pas sûr qu'on puisse en tirer des conclusions solides : il y a dix ans, IPv6 était peu présent dans la DFZ.

Ça, c'était le pourcentage des préfixes. Le nombre de préfixes à gérer a un effet négatif sur le routeur, car davantage de préfixes veut dire davantage de mémoire consommée par le routeur. La DFZ a actuellement plus de 700 000 préfixes (IPv4 et IPv6 mêlés). Une table de seulement 700 000 entrées ? Cela peut sembler peu à l'ère de la vidéo de chats en haute définition et du "*big data*". Certes, un routeur du cœur de l'Internet n'a pas la même architecture qu'un PC de bureau, et la mémoire n'y est pas disponible dans les mêmes quantités mais, quand même, est-ce si grave ?

En fait, le plus gênant pour le routeur typique n'est pas la quantité de préfixes mais le rythme de changement (le "*churn*"). Davantage de préfixes veut surtout dire davantage de changements à traiter. Huston regarde donc dans la suite de son article l'effet des changements. Il constate que les préfixes plus spécifiques sont un peu plus « bruyants » (davantage de changements), au moins en IPv4. Les préfixes plus spécifiques du deuxième cas (ingénierie de trafic) sont les plus bruyants, ce qui est assez logique. Attention avec ces statistiques, toutefois : Huston rappelle que le rythme de changement varie énormément selon les préfixes. Certains changent tout le temps, d'autres sont très stables.

Maintenant, venons-en aux actions possibles. Est-ce que ces préfixes plus spécifiques, émis par des opérateurs égoïstes et payés par tous, doivent être activement combattus, pour en réduire le nombre ? C'est l'idée qui est derrière des actions comme le "*CIDR report*" <`https://www.cidr-report.org/`>. En analysant automatiquement les préfixes de la DFZ, et en publiant la liste des opérateurs qui abusent le plus, on espère leur faire honte (cette technique du "*name and shame*" est très courante sur Internet, puisqu'il n'existe pas d'État central qui puisse faire respecter des règles et que les opérateurs n'ont, heureusement, pas de police ou d'armée privée), et diminuer avec le temps le pourcentage de préfixes « inutiles ». Par exemple, Huston calcule que la suppression de tous les plus spécifiques du cas 3 (recouvrement complet) diminuerait la table de routage globale de 20 % et le rythme des changements de 10 à 15 %.

En conclusion, Huston estime qu'on ne peut pas parler de tous les préfixes moins spécifiques de la même façon, certains sont réellement utiles. Certes, d'autres ne le sont pas, mais les supprimer serait une tâche longue et difficile, pour un bénéfice limité. Huston résume en disant que les préfixes plus spécifiques sont un problème agaçant, mais pas grave.

`https://www.bortzmeyer.org/bgp-more-specifics-ok.html`