RFC 6769 : Simple Virtual Aggregation (S-VA)

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 25 octobre 2012

Date de publication du RFC : Octobre 2012

https://www.bortzmeyer.org/6769.html

Un des gros problèmes des routeurs de la DFZ, dans l'Internet d'aujourd'hui, est l'augmentation du nombre de routes. Il y a, en octobre 2012, environ 450 000 <http://bgp.potaroo.net/> routes dans cette DFZ et ce nombre croît sans cesse. Ces routes doivent être mises en mémoire et la mémoire d'un routeur du cœur de l'Internet est quelque chose de compliqué, qu'on ne veut pas gaspiller. Des tas d'approches ont été proposées pour traiter ce problème. Ce RFC propose un moyen astucieux de limiter les dégâts, en permettant au routeur d'**agréger** les routes pour les préfixes IP proches. Ce RFC est donc très court car l'idée est simple.

Attention au terme de « table de routage », trop vague. Un routeur met des routes dans au moins deux tables, la RIB ("*Routing Information Base*", la base constituée par les protocoles de routage, BGP et ses copains OSPF et IS-IS) et la FIB ("*Forwarding Information Base*", typiquement placée dans la CAM car elle doit être très rapide, elle est consultée à chaque paquet). Actuellement, le routeur typique met toutes les entrées de sa RIB dans sa FIB. C'est sous-optimal car la RIB a beaucoup plus d'entrées que le routeur n'a de voisins réseau (des centaines de milliers de routes contre parfois seulement dix ou vingt routeurs voisins). Comme la taille de la FIB est souvent un facteur limitant (ce n'est pas le seul, donc ce RFC ne résout pas tous les problèmes de taille), il serait intéressant d'agréger violemment, en découplant RIB et FIB. Par exemple, si la RIB contient une entrée pour `2001:db8:16::/48` et une autre pour `2001:db8:17::/48`, et que les deux vont vers le même voisin (ont le même "*next hop*"), mettre les deux entrées dans la FIB est du gaspillage, une seule entrée, pour `2001:db8:16::/47`, suffirait. (Et le calcul pour déterminer ce qu'on peut agréger est simple.)

Au passage, on peut agréger dans la FIB mais moins dans la RIB car cette dernière doit être transmise aux routeurs pairs avec qui on échange des routes. Si on agrège, c'est une « compression avec perte » et les pairs ne recevront donc pas toute l'information. Néanmoins, ce RFC permet également de ne pas installer de routes dans la RIB. Celle-ci est typiquement dans une mémoire moins coûteuse mais on souhaite diminuer, non seulement la consommation mémoire mais également le rythme de changement. Moins on aura de préfixes, moins on aura de changements à gérer.

L'avantage de l'approche naïve ci-dessus est qu'elle est déployable unilatéralement : chaque routeur peut le faire dès aujourd'hui sans rien demander à personne. Son principal inconvénient est que la RIB n'est pas statique : si BGP change le "*next hop*" (le routeur suivant pour acheminer le paquet) de `2001:db8:17::/48`, il faut désagréger en vitesse et mettre deux entrées dans la FIB.

L'approche de ce RFC est donc un peu différente : certains routeurs de l'AS vont annoncer des **préfixes virtuels** ("*VA prefix*" pour "*Virtual Aggregation prefix*"). Ils seront transmis, à l'intérieur de l'AS uniquement, par les protocoles de routage habituels et tous les routeurs vont l'avoir dans leur RIB. Ces préfixes virtuels couvrent plusieurs préfixes « réels ». Les routeurs qui veulent économiser de la mémoire ne vont **pas** installer les préfixes réels couverts dans leur FIB. Notez bien que les préfixes virtuels ne sont pas annoncés en dehors de l'AS. Si cette solution n'est plus déployable par un routeur seul, elle l'est par un AS seul.

Un opérateur réseau va donc typiquement avoir quelques très gros routeurs qui auront dans leur FIB (et dans leur RIB) tous les préfixes. Ils annonceront en plus les préfixes virtuels. Les autres routeurs, moins gros et moins coûteux, garderont tous les préfixes uniquement dans leur RIB et ne mettront pas dans la FIB les préfixes qui sont couverts par un des préfixes virtuels. Le préfixe virtuel pourra même être la route par défaut, entraînant ainsi la non-installation dans la FIB de tous les préfixes ayant le même "*next hop*" que la route par défaut.

Un peu de vocabulaire, maintenant (section 1.1) : un **FIR** ("*FIB Installation Router*") est un des gros routeurs de cœur, qui installe toutes les routes dans sa FIB, et annonce les préfixes virtuels. Un **FSR** ("*FIR Suppressing Router*") est un routeur situé plus aux marges du réseau et qui n'installe pas dans sa FIB les routes couvertes par un préfixe virtuel, si elles ont le même "*next hop*" que le préfixe virtuel. (Le RFC dit « supprimer une route » pour « ne pas l'installer dans la FIB ».) Une troisième catégorie de routeurs est composé des machines actuelles, qui ne connaissent pas ce RFC, et ne peuvent donc ni émettre les préfixes virtuels, ni supprimer des routes couvertes par un tel préfixe.

La section 2 décrit en détail le fonctionnement de ce système. Un FIR doit annoncer les préfixes virtuels en mettant dans les attributs de l'annonce BGP un `ORIGIN` (RFC 4271 [1], section 5.1.1) à `INCOMPLETE` (pour dire que la route n'a pas été apprise par un protocole de routage). Il doit mettre `NEXT_HOP` (RFC 4271, section 5.1.3) à l'adresse du routeur qui gérera toutes les routes du préfixe virtuel. Et il doit ajouter un attribut `NO_EXPORT` (RFC 1997) pour éviter que ce préfixe virtuel ne sorte de l'AS.

Le FSR, lui, lorsqu'il reçoit de telles routes, les installe dans RIB et FIB. Il peut alors décider de supprimer de sa FIB les routes plus spécifiques ayant le même "*next hop*". Comme vous le voyez, c'est très simple.

La section 3 du RFC se penche sur les problèmes opérationnels pratiques. D'abord, attention au fait qu'une route BGP peut être utilisée pour résoudre un "*next hop*" qui n'est pas directement connecté. Il ne faudra donc pas supprimer de la RIB de telles routes.

On l'a vu, les routeurs FSR peuvent économiser de la place non seulement en FIB mais également en RIB, en supprimant de celles-ci les routes couvertes par un préfixe virtuel. Attention, cela peut interagir avec une politique de redistribution de routes aux pairs. Si une telle politique est en place (fondée, par exemple, sur la longueur du préfixe) on peut avoir des cas où la route couverte aurait été redistribuée

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

mais où le préfixe virtuel ne l'est pas. Dans ce cas, il ne faut pas supprimer la route couverte de la RIB (on peut toujours la supprimer de la FIB).

À noter que le groupe de travail grow <http://tools.ietf.org/wg/grow/> avait également travaillé sur des mécanismes d'agrégation plus complexes (avec des tunnels MPLS pour joindre les routeurs annonçant les préfixes virtuels) mais que cela n'a pas abouti. Les "*Internet-Drafts*" ont été abandonnés. Un bon article de fond sur cette idée d'« agrégation virtuelle » est celui de Paul Francis et Xiaohu Xu (un des auteurs du RFC), « "*Extending Router Lifetime with Virtual Aggregation*" <http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_13-1/131_aggregation.html> ».

Globalement, cela me semble une bonne idée. Avec les préfixes virtuels, ce n'est plus une solution locale au routeur mais c'est local à l'AS et donc ça reste déployable unilatéralement.

Merci à Sarah Nataf pour avoir attiré mon attention sur ce document et à Bastien Pilat et Benjamin Abadie pour leurs corrections.