RFC 7947 : Internet Exchange BGP Route Server

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 9 septembre 2016

Date de publication du RFC : Septembre 2016

https://www.bortzmeyer.org/7947.html

Ce nouveau RFC spécifie le comportement des serveurs de routes des points d'échange Internet. Un serveur de route collecte avec BGP les routes envoyées par ses pairs et les redistribue. Cela nécessite quelques ajustements du protocole BGP, expliqués ici. Un autre RFC, le RFC 7948 [1], décrit le côté opérationnel.

Un point d'échange Internet connecte un certain nombre d'acteurs (opérateurs, hébergeurs, FAI, etc) à un réseau commun, en général au niveau de la couche 2, et avec Ethernet. Chacun de ces acteurs gère son ou ses routeurs, connecté au[Caractère Unicode non montré [2]]x commutateur[Caractère Unicode non montré]s commun[Caractère Unicode non montré]s. Pour savoir à quel pair envoyer des paquets IP, ces routeurs ont besoin de s'informer mutuellement des préfixes IP qu'ils gèrent. Cela se fait avec le protocole BGP (RFC 4271).

Si le point d'échange rassemble N acteurs, connecter tout le monde nécessiterait N*(N-1)/2 sessions BGP. À chaque fois qu'un participant arriverait, il faudrait qu'il négocie N-1 accords de "*peering*". Et, en fonctionnement quotidien, il devrait superviser N-1 sessions BGP. L'idée de base du serveur de routes est d'introduire un intermédiaire (le serveur de routes, "*route server*"), qui reçoit les routes de chacun et les redistribue à tous. Ainsi, il n'y a qu'une seule session BGP à établir, avec le serveur de routes. Il est en général géré par l'opérateur du point d'échange. (Sur la plupart des points d'échange, l'usage de ce serveur de routes n'est pas obligatoire, et il y a donc également des sessions BGP directes entre les participants, pour des raisons variées.)

Le serveur de routes parle BGP avec ses pairs (les routeurs des participants au point d'échange) mais lui-même n'est pas un routeur : il ne transmet pas les paquets IP. Il travaille sur le plan de contrôle, pas sur celui des données.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7948.txt
2. Car trop difficile à faire afficher par LaTeX

Un serveur de routes est donc très proche d'un réflecteur (RFC 4456). La différence est que le réflecteur de routes travaille à l'intérieur d'un AS (protocole iBGP), alors que le serveur de routes travaille entre AS différents (protocole eBGP).

Un serveur de routes ressemble aussi à un collecteur de routes, comme ceux du RIS <https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris>. La différence est que le collecteur ne redistribue pas d'information à ses pairs BGP, il collecte des données, pour analyse, affichage, etc.

La section 2 du RFC forme le cœur de ce document. Elle rassemble les modifications du protocole BGP qui sont nécessaires pour le bon fonctionnement du serveur de routes. D'abord, le serveur ne doit pas modifier sans bonne raison les annonces qu'il reçoit : son but est de distribuer de l'information, pas de la bricoler. Le RFC demande ainsi que les attributs BGP bien connus (comme `AS_PATH`) ne soient pas modifiés, sauf très bonne raison. Ainsi, l'attribut `NEXT_HOP` qui indique l'adresse IP du routeur à qui faire suivre les paquets, ne doit pas être changé (ce que demandait la section 5.1.3 du RFC 4271) : le serveur de routes n'est pas un routeur, c'est à l'annonceur original qu'il faut transmettre les paquets. Idem pour `AS_PATH` cité plus haut : le serveur de routes ne doit pas mettre son propre numéro d'AS dans le chemin. De même, `MULTI_EXIT_DISC` (qui sert à choisir entre plusieurs routeurs d'un même AS voisin) doit être propagé aux clients du serveur de routes (normalement, il n'est pas propagé aux autres AS). Et, enfin, les communautés BGP (RFC 1997) indiquées dans une annonce ne doivent pas être changées, sauf évidemment si elles sont destinées au serveur de routes lui-même. On trouve des exemples de communautés destinées au serveur de routes dans la documentation du serveur de routes d'AMS-IX <https://ams-ix.net/technical/specifications-descriptions/ams-ix-route-servers> ou bien dans celle de celui de Netnod <http://www.netnod.se/ix/routeservers>, alors que celle de France-IX se trouve dans un objet RIPE <https://apps.db.ripe.net/search/query.html?full_query_string=&searchtext=AS51706&do_search=Search&inverse_attributes=None&ip_search_lvl=&alt_database=RIPE&object_type=aut-num&object_template=none&recursive=ON#resultsAnchor>. Il y a aussi des serveurs de routes (comme ceux du France-IX qui étiquettent les annonces apprises avec une communauté indiquant où la route a été apprise (dans quel POP.)

Par défaut, toutes les routes de tous les clients sont distribuées à tous les autres clients. C'est le but d'un serveur de routes. Mais on peut souhaiter parfois une certaine restriction à cette redistribution. Cela peut être mis en œuvre par le serveur de routes (puisque le routeur original de l'annonce parle au serveur de routes, pas au client final). Idéalement, le serveur de routes devrait donc maintenir une base de routes, et un processus de décision BGP, pour chacun de ses clients. Cela permet de coller au mode de fonctionnement normal de BGP. Et cela ne nécessite aucun changement chez les clients. Mais c'est plus coûteux en ressources pour le serveur de routes. (Pour BIRD, voir l'option `secondary`, qui réduit l[Caractère Unicode non montré]usage mémoire et CPU de manière importante, expliquée dans cet exposé <https://www.nanog.org/meetings/nanog57/presentations/Wednesday/wed.general.Filip.BIRD.16.pdf>.)

Les autres solutions à ce problème d'un filtrage par client nécessitent des extensions à BGP, comme le "*Diverse path*" du RFC 6774 ou comme la future option `ADD-PATH`.

L'annexe de ce RFC consacrée aux remerciements résume un peu d'histoire des serveurs de routes. La première description de ce système date de 1995, dans le RFC 1863 (depuis reclassé comme « d'intérêt historique seulement », cf. RFC 4223).

Plusieurs mises en œuvre de serveurs de route existent évidemment, conformes à ce RFC. C'est par exemple le cas chez Cisco, BIRD et Quagga (voir le rapport technique sur ces programmes <http://trac.tools.ietf.org/wg/idr/trac/wiki/draft-ietf-idr-ix-bgp-route-server%20implementati

Merci à Arnaud Fenioux pour sa relecture et ses ajouts et corrections.