RFC 8195 : Use of BGP Large Communities

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 1 juillet 2017

Date de publication du RFC : Juin 2017

https://www.bortzmeyer.org/8195.html

Le RFC 8092[1] a normalisé la notion de « grande communauté » BGP, des données attachées aux annonces de routes, plus grandes que les précédentes « communautés » et permettant donc de stocker davantage d'informations. Ce nouveau RFC explique à quoi peuvent servir ces grandes communautés et donne des exemples d'utilisation. Un document concret, qui ravira les opérateurs réseaux.

Le RFC original sur les communautés BGP était le RFC 1997, et il était accompagné d'un RFC 1998 présentant des cas concrets d'utilisation. C'est la même démarche qui est faite ici, de faire suivre le document de normalisation, le RFC 8092, d'un document décrivant des utilisations dans le monde réel (avec des expériences pratiques décrites à NANOG ou NLnog).

Un petit rappel des grandes communautés, d'abord (section 2 du RFC). Chaque communauté se compose de trois champs de quatre octets chacun. Le premier champ se nomme GA, pour "*Global Administrator*", et sa valeur est le numéro de l'AS qui a ajouté cette communauté (rappelez-vous que les numéros d'AS font désormais quatre octets), ce qui assure l'unicité mondiale des communautés. Les deux autres champs portent les noms peu imaginatifs de "*Local Data Part 1*" et "*Local Data Part 2*". La plupart du temps, on se sert du premier pour identifier une fonction, et du second pour identifier les paramètres d'une fonction. Notez qu'il n'existe pas de standard pour ces deux champs locaux : comme leur nom l'indique, chaque AS les affecte comme il veut. Une même valeur peut donc avoir des significations différentes selon l'AS.

Le RFC contient plusieurs exemples, qui utilisent à chaque fois la même topologie : un transitaire, l'AS 65551, avec un client, l'AS 64497, qui est lui-même transitaire pour les AS 64498 et 64499, qui par ailleurs ont une relation de "*peering*" (en épais sur l'image) :

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc8092.txt

Notre RFC classe les communautés en deux catégories : celles d'**information** et celles d'**action**. Les premières servent à distribuer des informations qui ne modifieront pas le comportement des routeurs, mais pourront être utiles pour le débogage, ou les statistiques. Un exemple est l'ajout d'une communauté pour indiquer dans quel pays on a appris telle route. Elles sont typiquement ajoutées par l'AS qui a défini ces informations, et il mettra donc son numéro d'AS dans le champ GA. Les secondes, les communautés d'action, servent à indiquer une action souhaitée. Elles sont définies par l'AS qui aura à agir, et mises dans le message BGP par un de ses pairs. Un exemple est une communauté pour indiquer à son pair quelle préférence on souhaite qu'il attribue à telle route.

La section 3 donne des exemples de communautés d'**information**. Le premier est celui où l'AS 66497 marque les routes qu'il a reçues avec une communauté qui indique le pays où la route a été apprise (le code pays utilisé est celui de ISO 3166-1). La fonction a reçu le numéro 1. Donc, la présence dans l'annonce BGP de la communauté `64497:1:528` indique une route apprise aux Pays-Bas, la communauté `64497:1:392` indique le Japon, etc. Le même opérateur peut aussi prévoir une fonction 2 pour indiquer la région (concept plus vaste que le pays), en utilisant les codes M.49 : `64497:2:2` dit que la route a été apprise en Afrique, `64497:2:150` en Europe et ainsi de suite. Rappelez-vous bien que la signification d'une fonction (et donc des paramètres qui suivent) **dépend de l'AS** (ou, plus rigoureusement, du champ GA). Il n'y a pas de standardisation des fonctions. Si vous voyez une communauté `65551:2:2`, cela ne signifie pas forcément que la route vient d'Afrique : l'AS 65551 peut utiliser la fonction 2 pour tout à fait autre chose.

À part l'origine géographique d'une route, il est souvent utile de savoir si la route a été apprise d'un pair ou d'un transitaire. Le RFC donne l'exemple d'une fonction 3 où le paramètre 1 indique une route interne, 2 une route apprise d'un client, 3 d'un pair et 4 d'un transitaire. Ainsi, la communauté `64497:3:2` nous dit que la route vient d'un client de l'AS 64497.

Une même annonce de route peut avoir plusieurs communautés. Une route étiquetée `64497:1:528`, `64497:2:150` et `64497:3:3` vient donc d'un pair aux Pays-Bas.

Et les communautés d'**action** (section 4 du RFC) ? Un premier exemple est celui où on indique à son camarade BGP de ne pas exporter inconditionnellement une route qu'on lui annonce (ce qui est le comportement par défaut de BGP). Mettons que c'est la fonction 4. (Vous noterez qu'il n'existe pas d'espace de numérotation de fonctions distinct pour les communautés d'information et d'action. On sait que la fonction 4 est d'action uniquement si on connait la politique de l'AS en question.) Mettons encore que l'AS 64497 a défini le paramètre comme étant un numéro d'AS à qui il ne faut **pas** exporter la route. Ainsi, envoyer à l'AS 64497 une route étiquetée `64497:4:64498` signifierait « n'exporte **pas** cette route à l'AS 64498 ». De la même façon, on peut imaginer une fonction 5 qui utilise pour cette non-exportation sélective le code pays. `64997:5:392` voudrait dire « n'exporte pas cette route au Japon ». (Rappelez-vous que BGP est du routage politique : son but principal est de permettre de faire respecter les règles du business des opérateurs.)

Je me permets d'enfoncer le clou : dans les communautés d'action, l'AS qui ajoute une communauté ne met pas **son** numéro d'AS dans le champ GA, mais celui de l'AS qui doit agir.

Un autre exemple de communauté d'action serait un allongement sélectif du chemin d'AS. On allonge le chemin d'AS en répétant le même numéro d'AS lorsqu'on veut décourager l'utilisation d'une certaine route. (En effet, un des facteurs essentiels d'une décision BGP est la longueur du chemin d'AS : BGP préfère le plus court.) Ainsi, l'exemple donné utilise la fonction 6 avec comme paramètre l'AS voisin où appliquer cet allongement ("*prepending*"). `64497:6:64498` signifie « ajoute ton AS quand tu exportes vers 64498 ».

Quand on gère un routeur BGP multihomé, influencer le trafic sortant est assez simple. On peut par exemple mettre une préférence plus élevée à la sortie vers tel transitaire. Influencer le trafic entrant (par exemple si on veut que le trafic vienne surtout par tel transitaire) est plus délicat : on ne peut pas configurer directement la politique des autres AS. Certains partenaires BGP peuvent permettre de définir la préférence locale (RFC 4271, section 5.1.5), via une communauté. Ici, le RFC donne l'exemple d'une fonction 8 qui indique « mets la préférence d'une route client [a priori une préférence élevée] », 10 une route de "*peering*", 11 de transit et 12 une route de secours, à n'utiliser qu'en dernier recours (peut-être parce qu'elle passe par un fournisseur cher, ou bien de mauvaise qualité). Le paramètre (le troisième champ de la communauté) n'est pas utilisé. Alors, la communauté `64997:12:0` signifiera « AS 64997, mets à cette route la préférence des routes de secours [a priori très basse] ».

Le RFC suggère aussi, si le paramètre est utilisé, de le prendre comme indiquant la région où cette préférence spécifique est appliquée. Par exemple `64997:10:5` demandera à l'AS 64997 de mettre la préférence des routes de "*peering*" (10) à tous les routeurs en Amérique du Sud (code 5 de la norme M.49).

Attention, changer la préférence locale est une arme puissante, et sa mauvaise utilisation peut mener à des coincements (RFC 4264).

Dernier exemple donné par notre RFC, les souhaits exprimés à un serveur de routes (RFC 7947). Ces serveurs parlent en BGP mais ne sont pas des routeurs, leur seul rôle est de redistribuer l'information aux routeurs. Par défaut, tout est redistribué à tout le monde (RFC 7948), ce qui n'est pas toujours ce qu'on souhaite. Si on veut une politique plus sélective, certains serveurs de route documentent des communautés que les clients du serveur peuvent utiliser pour influencer la redistribution des routes. Notre RFC donne comme exemple une fonction 13 pour dire « n'annonce pas » (alors que cela aurait été le comportement normal) et une fonction 14 pour « annonce quand même » (si le comportement normal aurait été de ne pas relayer la route). Ainsi, `64997:13:0` est « n'annonce pas cette route par défaut » et `64997:14:64999` est « annonce cette route à l'AS 64999 ».

Vous voudriez des exemples réels ? Netnod avait annoncé <`https://www.netnod.se/sites/default/files/2017-04/Mattias_Karlsson_Netnod_ix_technichal_update_spring_meeting_2017_1.pdf`> que ces grandes communautés étaient utilisables sur leur AS 52005 :

— `52005:0:0` "*Do not announce to any peer*"
— `52005:0:ASx` "*Do not announce to ASx*"
— `52005:1:ASx` "*Announce to ASx if 52005 :0 :0 is set*"
— `52005:101:ASx` "*Prepend peer AS 1 time to ASx*"
— `52005:102:ASx` "*Prepend peer AS 2 times to ASx*"
— `52005:103:ASx` "*Prepend peer AS 2 times to ASx*"

On trouve également ces communautés au IX-Denver <`http://ix-denver.org/`>, documentées ici <`http://ix-denver.org/?q=technical`> (sous « "*Route Server B and C Communities*" ») ou au SIX <`https://www.seattleix.net/`> (dans leur documentation <`https://www.seattleix.net/route-servers#communities`>). Il y a enfin cet exemple <`http://largebgpcommunities.net/2016/ecix-embraces-large-communities/`> à ECIX. (Merci à Job Snijders et Pier Carlo Chiodi[Caractère Unicode non montré [2]] pour les exemples.)

Voilà, à vous maintenant de développer votre propre politique de communautés d'information, et de lire la documentation de vos voisins BGP pour voir quelles communautés ils utilisent (communautés d'action).

2. Car trop difficile à faire afficher par LaTeX