RFC 1997 : BGP Communities Attribute

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 17 février 2017

Date de publication du RFC : Août 1996

https://www.bortzmeyer.org/1997.html

Il n'y a pas de rapport simple entre la longueur d'un RFC et l'importance de la norme technique qu'il spécifie. Ce RFC (vieux de plus de vingt ans) ne fait que cinq pages mais décrit une technique qui est essentielle au bon fonctionnement du routage sur l'Internet, la technique des **communautés** BGP.

Rien de plus simple que cette idée : une **communauté** est juste une étiquette numérique qu'on ajoute aux annonces de routes IP transportées par le protocole BGP. On peut ainsi « décorer » ses annonces comme on veut, et les autres routeurs BGP pourront ainsi prendre des décisions fondées sur cette information supplémentaire. Les communautés sont juste une syntaxe, on leur met la signification qu'on veut.

Un tout petit rappel sur BGP : c'est **le** protocole d'échange de routes entre les opérateurs Internet. Son principe est simple (RFC 4271 [1]) : quand une route apparait, on annonce à ses pairs BGP la route, sous la forme d'un préfixe d'adresses IP, avec un certain nombre d'**attributs**. Les attributs ont un format et une sémantique précise. Voici un exemple d'une annonce reçue par le service RouteViews <http://www.routeviews.org/>, et affichée sous forme texte par le logiciel bgpdump <https://bitbucket.org/ripencc/bgpdump/wiki/Home> :

```
TIME: 02/17/17 15:00:00
TYPE: BGP4MP/MESSAGE/Update
FROM: 208.51.134.246 AS3549
TO: 128.223.51.102 AS6447
ORIGIN: IGP
ASPATH: 3549 3356 2914 30259
NEXT_HOP: 208.51.134.246
MULTI_EXIT_DISC: 13920
ATOMIC_AGGREGATE
AGGREGATOR: AS30259 10.11.1.1
COMMUNITY: 2914:410 2914:1001 2914:2000 2914:3000 3356:3 3356:86 3356:575 3356:666 3356:2011 3356:11940 3549:201
WITHDRAW
  93.181.192.0/19
ANNOUNCE
  199.193.160.0/22
```

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

On y voit que le routeur 208.51.134.246, appartenant à l'AS 3549 (Level 3, ex-Global Crossing) a annoncé une route à destination du préfixe 199.193.160.0/22 (il a aussi retiré une autre route mais on ne s'en soucie pas ici). Cette annonce avait plusieurs attributs comme le chemin d'AS (ASPATH) emprunté. Les communautés (au nombre de 21 ici) sont un attribut COMMUNITY dont le format est défini mais dont on fait ensuite ce qu'on veut. L'utilisation la plus courante est d'indiquer l'origine d'une route, pour d'éventuelles décisions ultérieures, en fonction de la politique de routage. Les communautés sont donc un outil pour gérer la complexité de ces politiques.

Le RFC définit d'ailleurs une communauté comme « un groupe de destinations partageant une propriété commune ». Ainsi, dans le cas de l'annonce ci-dessus, la lecture de la documentation des différents opérateurs nous apprend que 3549:31826 indique <https://onestep.net/communities/as3549/> que la route a été apprise en Europe, au Royaume-Uni, que 2914:410 nous montre <https://onestep.net/communities/as2914/> qu'il s'agissait d'une route d'un client (et non pas d'un pair) de NTT, etc.

L'exemple d'utilisation donné par le RFC date pas mal (NSFNET n'existe plus) mais ce genre de cas est toujours fréquent. NSFNET, financé par l'argent public, ne permettait pas d'utilisation purement commerciale. Une entreprise à but lucratif pouvait s'y connecter, mais seulement pour échanger avec les organismes de recherche ou d'enseignement (le RFC parle d'organismes respectant l'AUP, qui étaient les conditions d'utilisation de NSFNET). Une telle politique est facile à faire avec les communautés : on étiquette toutes les routes issues du monde enseignement/recherche avec une communauté signifiant « route AUP », et NSFNET pouvait annoncer les routes AUP à tous et les routes non-AUP (n'ayant pas cette communauté) seulement aux client AUP. Ainsi, deux entreprises commerciales ne pouvaient pas utiliser NSFNET pour communiquer entre elles. Sans les communautés, une telle politique aurait nécessité une base complexe de préfixes IP, base difficile à maintenir, d'autant plus que tous les routeurs de bord devaient y accéder. (Avant les communautés, c'était bien ainsi qu'on procédait, avec les retards et les erreurs qu'on imagine.)

Autre exemple d'utilisation donné par le RFC, l'agrégation de routes. Si on annonce à la fois un préfixe englobant et un sous-préfixe plus spécifique pour optimiser l'accès à un site particulier, on ne souhaite en général annoncer ce sous-préfixe qu'aux pairs proches (les autres n'ont pas de chemin meilleur vers ce site). On va donc étiqueter l'annonce faite à ces pairs proches avec une communauté indiquant « cette route est pour vous, mais ne la propagez pas ». D'autres exemples d'utilisation figurent dans les RFC 1998 et RFC 4384.

L'attribut COMMUNITY (le RFC le nomme COMMUNITIES, ce qui est effectivement plus logique, mais il a bien été enregistré sous le nom COMMUNITY) est donc un attribut optionnel (certaines annonces BGP ne l'utiliseront pas) et transitif (c'est-à-dire qu'il est conçu pour être transmis avec l'annonce lorsqu'on la relaie à ses pairs). Il consiste en un ensemble (non ordonné, donc) de communautés, chacune occupant quatre octets (ce qui est bien insuffisant aujourd'hui). Son code de type d'attribut est 8. Le nombre de communautés dans une annonce est très variable. Par exemple, le LU-CIX voit une moyenne de 10,5 communautés par route sur ses serveurs de routes.

Si un attribut COMMUNITY est mal formé, en vertu du RFC 7606, la route annoncée sera retirée. (À l'époque du RFC originel, une erreur aboutissait à fermer toute la session BGP, retirant toutes les routes.)

Chaque communauté peut donc aller de 0x0000000 à 0xFFFFFFFF mais les valeurs de 0x0000000 à 0x0000FFFF, et de 0xFFFF0000 à 0xFFFFFFFF sont réservées. D'autre part, la convention recommandée est de mettre son numéro d'AS dans les deux premiers octets, et une valeur locale à l'AS dans les deux derniers. (Notez que ce système ne marche plus avec les AS de quatre octets du RFC 6793, ce qui a mené aux RFC 4360 et RFC 8092.) Prenons comme exemple de communauté 0x0D1C07D1. On note les communautés sous forme de deux groupes de deux octets chacun, séparés par un deux-points. Cette

communauté est donc `3356:2001` : AS 3356 (Level 3) et valeur locale 2001 (le choix des valeurs locales est une décision... locale donc on ne peut savoir ce que signifie 2001 qu'en regardant la documentation de Level 3 <https://apps.db.ripe.net/search/lookup.html?source=ripe&key=AS3356&type=aut-num>. Dit autrement, la valeur locale est opaque.)

Certaines valeurs sont réservées à des communautés <https://www.iana.org/assignments/bgp-well-known-communities/bgp-well-known-communities.xml> « bien connues ». C'est le cas par exemple de 0xFFFFFF01 (alias `NO_EXPORT` : ne pas transmettre cette annonce en dehors de son AS), de 0xFFFFFF02 (`NO_ADVERTISE`, ne transmettre cette annonce à aucun autre routeur) ou bien de la plus récente 0xFFFF029A (`BLACKHOLE`, RFC 7999). Rappelez-vous que chaque routeur est maître de ses décisions : les communautés bien connues sont une suggestion, mais on ne peut jamais être sûr que le pair va la suivre (c'est ainsi que, malgré les `NO_EXPORT` que mettent les nœuds "*anycast*" qui veulent rester relativement locaux, on voit dans certains cas les annonces se propager plus loin, parfois pour des bonnes et parfois pour des mauvaises raisons.)

Enfin, le RFC précise qu'un routeur est libre d'ajouter ses propres communautés aux annonces qu'il relaie, voire de supprimer des communautés existantes (chacun est maître de son routage).

Quelques bonnes lectures sur les communautés BGP :

— Un gros effort de rassemblement des documentations de tous les opérateurs sur leurs communautés <http://www.onesc.net/communities/>. Si vous voulez savoir ce que signifie `8308:60050`, c'est là qu'il faut aller (réponse ici <https://onestep.net/communities/as8308/>).
— Un très bon cours du NSRC <https://nsrc.org/workshops/2016/apricot2016/raw-attachment/wiki/Track2BGP/09-BGP-Communities.pdf>, très concret, avec plein de détails (avec exemples de configuration pour Cisco IOS).
— Toujours pour les possesseurs de Cisco, la documentation officielle <http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/28784-bgp-community.html>.
— Un bon exposé à NANOG <https://www.nanog.org/meetings/nanog40/presentations/BGPcommunities.pdf>.
— Un excellent travail de recherche et de taxonomie sur les communautés <http://inl.info.ucl.ac.be/communities>.
— Un autre travail de recherche, pour extraire de l'information sur les clients des opérateurs <https://labs.ripe.net/Members/emileaben/a-tale-of-bgp-collectors-and-customer-cones>, en utilisant leurs communautés (notez celles d'AT&T, bien détaillées).
— Un exemple de politique de communautés extrêmement complète, celle de NTT <http://www.us.ntt.net/support/policy/routing.cfm>. Notez le "*European country origins*" (là où la route a été apprise), qui peut permettre de faire du routage Schengen.
— Une communauté bien connue qui a besoin d'une bonne documentation est la `BLACKHOLE` du RFC 7999. Voyez par exemple celle de Hurricane Electric <https://www.he.net/adm/blackhole.html>.

Certains "*looking glass*" affichent les communautés par exemple celui de Cogent <http://www.cogentco.com/en/network/looking-glass> :

```
 BGP routing table entry for 129.250.0.0/16, version 3444371605
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  2914
    130.117.14.250 (metric 10109031) from 38.28.1.83 (38.28.1.83)
      Origin IGP, metric 4294967294, localpref 100, valid, internal, best
      Community: 174:11102 174:20666 174:21100 174:22010
      Originator: 38.28.1.32, Cluster list: 38.28.1.83, 38.28.1.67, 38.28.1.235
```

Les communautés sont souvent documentées dans l'objet AS stocké dans la base d'un RIR et accessible via whois (ou, aujourd'hui, RDAP). Ici, celle du France-IX (notez l'utilisation d'AS privés) :

```
% whois AS51706
...
aut-num:        AS51706
as-name:        FRANCE-IX-AS
...
remarks:        The following communities can be used by members:
remarks:
remarks:        ****************************************************
remarks:        ** Note: These communities are evaluated
remarks:        ** on a "first match win" basis
remarks:        ****************************************************
remarks:        0:peer-as = Don't send route to this peer as
remarks:        51706:peer-as = Send route to this peer as
remarks:        0:51706 = Don't send route to any peer
remarks:        51706:51706 = Send route to all peers
remarks:        ****************************************************
remarks:        ** Note: the community (51706:51706) is applied
remarks:        ** by default by the route-server
remarks:        ****************************************************
remarks:
remarks:        65101:peer-as = Prepend 1x to this peer
remarks:        65102:peer-as = Prepend 2x to this peer
remarks:        65103:peer-as = Prepend 3x to this peer
remarks:        65201:peer-as = Set MED 50 to this peer
remarks:        65202:peer-as = Set MED 100 to this peer
remarks:        65203:peer-as = Set MED 200 to this peer
remarks:
remarks:        ----------------------------------------------------
remarks:        BLACKHOLING, set the next-hop to the blackhole router
remarks:        can be use with the basic community (above)
remarks:
remarks:        65535:666 = BLACKHOLE [RFC7999]
remarks:
remarks:        https://www.franceix.net/en/technical/blackholing/
remarks:
remarks:        ----------------------------------------------------
remarks:        Set peer-as value as listed below for all IXP members:
remarks:        (Can't be used for 51706:peer-as)
remarks:        64649 = FranceIX Marseille peers
remarks:        64650 = FranceIX Paris peers
remarks:        64651 = SFINX peers
remarks:        64652 = LyonIX peers
remarks:        64653 = LU-CIX peers
remarks:        64654 = TOP-IX peers
remarks:        64655 = TOUIX peers
remarks:
remarks:        ----------------------------------------------------
remarks:        Set peer-as value as listed below for 32 bits ASNs:
remarks:        AS197422 -> AS64701 (Tetaneutral)
remarks:        AS196689 -> AS64702 (Digicube)
[...]
remarks:
remarks:        Extended Communities are supported and usage is
remarks:        encouraged instead of 32b->16b mapping
remarks:        ----------------------------------------------------
remarks:        Communities that are in the public range
remarks:        (1-64495:x) and (131072-4199999999:x)
remarks:        will be preserved by the route-servers
remarks:        ----------------------------------------------------
remarks:        Well-known communities are not interpreted by the
remarks:        route-servers and are propagated to all peers
remarks:        ----------------------------------------------------
remarks:
remarks:        The following communities are applied by the route-server:
remarks:
remarks:        ****************************************************
remarks:        ** WARNING
remarks:        ** You should not set any of these by yourself
remarks:        ** (from 51706:64495 to 51706:64699)
```

```
remarks:        ** (and 51706:64800 to 51706:65535)
remarks:        ** If you do so, your routes will be rejected
remarks:        **************************************************
remarks:
remarks:        51706:64601 = Prefix received from a peer on RS1 Paris
remarks:        51706:64602 = Prefix received from a peer on RS2 Paris
remarks:        51706:64611 = Prefix received from a peer on RS1 Marseille
remarks:        51706:64612 = Prefix received from a peer on RS2 Marseille
remarks:
remarks:        51706:64649 = Prefix received from a FranceIX Marseille peer
remarks:        51706:64650 = Prefix received from a FranceIX Paris peer
remarks:        51706:64651 = Prefix received from a SFINX peer
remarks:        51706:64652 = Prefix received from a LyonIX peer
remarks:        51706:64653 = Prefix received from a LU-CIX peer
remarks:        51706:64654 = Prefix received from a TOP-IX peer
remarks:        51706:64655 = Prefix received from a TOUIX peer
remarks:
remarks:        51706:64666 = Prefix with invalid route origin
...
```