RFC 8326 : Graceful BGP Session Shutdown

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 7 mars 2018

Date de publication du RFC : Mars 2018

https://www.bortzmeyer.org/8326.html

Voici une nouvelle communauté BGP, GRACEFUL_SHUTDOWN, qui va permettre d'annoncer une route à son pair BGP, tout en l'avertissant que le lien par lequel elle passe va être bientôt coupé pour une maintenance prévue. Le pair pourra alors automatiquement limiter l'usage de cette route et chercher tout de suite des alternatives. Cela évitera les pertes de paquets qui se produisent quand on arrête un lien ou un routeur.

Le protocole BGP (RFC 4271 [1]) qui assure le routage entre les AS qui composent l'Internet permet aux routeurs d'échanger des informations entre eux « pour aller vers 2001:db8:bc9::/48, passe donc par moi ». Avec ces informations, chaque routeur calcule les routes à suivre pour chaque destination. Une fois que c'est fait, tout le monde se repose ? Non, parce qu'il y a tout le temps des changements. Certains peuvent être imprévus et accidentels (la fameuse pelleteuse <https://twitter.com/AlertePelleteuz>), d'autres sont planifiés à l'avance : ce sont les opérations de maintenance « le 7 février à 2300 UTC, nous allons remplacer une "*line card*" du routeur, coupant toutes les sessions BGP de cette carte ». Voici par exemple un message reçu sur la liste de diffusion des opérateurs connectés au France-IX :

```
Date: Wed, 31 Jan 2018 15:57:34 +0000
From: Quelqu'un <quelquun@opérateur>
To: "paris@members.franceix.net" <paris@members.franceix.net>
Subject: [FranceIX members] [Paris] [OPÉRATEUR/ASXXXXX] - France-IX port  maintenance

Dear peers,

Tomorrow morning CEST, we will be conducting a maintenance that will
impact one of our connection to France-IX (IPs: x.y.z.t/2001:x:y:z::t).

All sessions will be shut down before and brought back up once the maintenance will be over.

Please note that our MAC address will change as the link will be migrated to a new router.

Cheers
```

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

Lorsqu'une telle opération est effectuée, le résultat est le même que pour une coupure imprévue : les sessions BGP sont coupées, les routeurs retirent les routes apprises via ces sessions, et vont chercher d'autres routes dans les annonces qu'ils ont reçues. Ils propagent ensuite ces changements à leurs voisins, jusqu'à ce que tout l'Internet soit au courant. Le problème est que cela prend du temps (quelques secondes au moins, des dizaines de secondes, parfois, à moins que les routeurs n'utilisent le RFC 7911 mais ce n'est pas toujours le cas). Et pendant ce temps, les paquets continuent à arriver à des routeurs qui ne savent plus les traiter (section 3 du RFC). Ces paquets seront jetés, et devront être réémis (pour le cas de TCP). Ce n'est pas satisfaisant. Bien sûr, quand la coupure est imprévue, il n'y a pas le choix. Mais quand elle est planifiée, on devrait pouvoir faire mieux, avertir les routeurs qu'ils devraient cesser d'utiliser cette route. C'est justement ce que permet la nouvelle communauté `GRACEFUL_SHUTDOWN`. (Les communautés BGP sont décrites dans le RFC 1997.) Elle s'utilise **avant** la coupure, indiquant aux pairs qu'ils devraient commencer le recalcul des routes, mais qu'ils peuvent continuer à utiliser les anciennes routes pendant ce temps. (Notez qu'un cahier des charges avait été établi pour ce problème, le RFC 6198. Et que ce projet d'une communauté pour les arrêts planifiés est ancien, au moins dix ans.)

Ce RFC décrit donc deux choses, la nouvelle communauté normalisée, `GRACEFUL_SHUTDOWN` (section 5), et la procédure à utiliser pour s'en servir proprement (section 4). L'idée est que les routes qui vont bientôt être coupées pour maintenance restent utilisées, mais avec une préférence locale très faible (la valeur 0 est recommandée, la plus petite valeur possible). La notion de préférence locale est décrite dans le RFC 4271, section 5.1.5. Comme son nom l'indique, elle est locale à un AS, et représente sa préférence (décidée unilatéralement) pour une route plutôt que pour une autre. Lors du choix d'une route par BGP, c'est le premier critère consulté.

Pour mettre en œuvre cette idée, chaque routeur au bord des AS (ASBR, pour "*Autonomous System Border Router*") doit avoir une règle qui, lorsqu'une annonce de route arrive avec la communauté `GRACEFUL_SHUTDOWN`, applique une préférence locale de 0. Notez que cela peut se faire avec les routeurs actuels, aucun code nouveau n'est nécessaire, ce RFC ne décrit qu'une procédure. Une fois que cette règle est en place, tout le reste sera automatique, chez les pairs de l'AS qui coupe un lien ou un routeur.

Et l'AS qui procéde à une opération de maintenance, que doit-il faire ? Dans l'ordre :

— Sur la[Caractère Unicode non montré [2]]es session[Caractère Unicode non montré]s BGP qui va[Caractère Unicode non montré]ont être coupée[Caractère Unicode non montré]s, appliquer la communauté `GRACEFUL_SHUTDOWN` aux routes qu'on annonce ("*outbound policy*"),
— Sur la[Caractère Unicode non montré]es session[Caractère Unicode non montré]s BGP qui va[Caractère Unicode non montré]ont être coupée[Caractère Unicode non montré]s, appliquer la communauté `GRACEFUL_SHUTDOWN` aux routes qu'on reçoit ("*inbound policy*"), **et** mettre leur préférence locale à zéro,
— Attendre patiemment que tout le monde ait convergé (propagation des annonces),
— Procéder à l'opération de maintenance, qui va couper BGP (et c'est plus joli si on utilise le RFC 9003).

J'ai dit plus haut qu'il n'était pas nécessaire de modifier le logiciel des routeurs BGP mais évidemment tout est plus simple s'ils connaissent la communauté `GRACEFUL_SHUTDOWN` et simplifient ainsi la tâche de l'administrateur réseaux. Cette communauté est « bien connue » (elle n'est pas spécifique à un AS), décrite dans la section 5 du RFC, enregistrée à l'IANA `<https://www.iana.org/assignments/bgp-well-known-communities/bgp-well-known-communities.xml>` et sa valeur est 0xFFFF0000 (qui peut aussi s'écrire 65535 :0, dans la notation habituelle des communautés).

La section 6 du RFC fait le tour de la sécurité de ce système. Comme il permet d'influencer le routage chez les pairs (on annonce une route avec la communauté `GRACEFUL_SHUTDOWN` et paf, le pair met une

2. Car trop difficile à faire afficher par LaTeX

très faible préférence à ces routes), il ouvre la porte à de l'ingénierie du trafic pas toujours bienveillante. Il est donc prudent de regarder ce qu'annoncent ses pairs, et d'engueuler ou de dépairer ceux et celles qui abusent de ce mécanisme.

Pour les amateurs de solutions alternatives, l'annexe A explique les autres techniques qui auraient pu être utilisées lors de la réception des routes marquées avec `GRACEFUL_SHUTDOWN`. Au lieu d'influencer la préférence locale, on aurait par exemple pu utiliser le MED ("*multi-exit discriminator*", RFC 4271, section 5.1.4) mais il n'est considéré par les pairs qu'après d'autres critères, et il ne garantit donc pas que le lien bientôt coupé ne sera plus utilisé.

L'annexe B donne des exemples de configuration pour différents types de routeurs. (Configurations pour l'AS qui reçoit la notification d'un arrêt proche, pas pour ceux qui émettent.) Ainsi, pour IOS XR :

```
!    65535:0 = 0xFFFF0000
community-set comm-graceful-shutdown
  65535:0
end-set

route-policy AS64497-ebgp-inbound
  ! Règles appliquées aux annonces reçues du pair, l'AS 64497. Bien
  ! sûr, en vrai, il y aurait plein d'autres règles, par exemple de filtrage.
  if community matches-any comm-graceful-shutdown then
    set local-preference 0
  endif
  ! On a appliqué la règle du RFC : mettre la plus faible
  ! préférence locale possible
end-policy

! La configuration de la session BGP avec le pair
router bgp 64496
 neighbor 2001:db8:1:2::1
  remote-as 64497
  address-family ipv6 unicast
   send-community-ebgp
   route-policy AS64497-ebgp-inbound in
```

Pour BIRD, cela sera :

```
# (65535, 0) = 0xFFFF0000
function honor_graceful_shutdown() {
    if (65535, 0) ~ bgp_community then {
        bgp_local_pref = 0;
    }
}
filter AS64497_ebgp_inbound
{
        # Règles appliquées aux annonces reçues du pair, l'AS 64497. Bien
        # sûr, en vrai, il y aurait plein d'autres règles, par
# exemple de filtrage.
        honor_graceful_shutdown();
}
protocol bgp peer_64497_1 {
    neighbor 2001:db8:1:2::1 as 64497;
    local as 64496;
    import keep filtered;
    import filter AS64497_ebgp_inbound;
}
```

Et sur OpenBGPD (on voit qu'il connait `GRACEFUL_SHUTDOWN`, il n'y a pas besoin de donner sa valeur) :

```
AS 64496
router-id 192.0.2.1
neighbor 2001:db8:1:2::1 {
        remote-as 64497
}
# Règles appliquées aux annonces reçues du pair, l'AS 64497. Bien
# sûr, en vrai, il y aurait plein d'autres règles, par exemple de filtrage.
match from any community GRACEFUL_SHUTDOWN set { localpref 0 }
```

Enfin, l'annexe C du RFC décrit quelques détails supplémentaires, par exemple pour IBGP (BGP interne à un AS).

Notez que ce nouveau RFC est prévu pour le cas où la transmission des paquets ("*forwarding plane*") est affectée. Si c'est uniquement la session BGP ("*control plane*") qui est touchée, la solution du RFC 4724, "*Graceful Restart*", est plus appropriée.