Groupe de travail Réseau
Request for Comments : 4456
RFC rendues obsolètes : 2796, 1966
Catégorie : Sur la voie de la normalisation
Traduction Claude Brière de L'Isle

T. Bates, Cisco Systems
E. Chen, Cisco Systems
R. Chandra, Sonoa Systems
avril 2006

Réflexion de chemin BGP : une solution de remplacement au BGP interne à maillage complet (IBGP)

Résumé
Le protocole de routeur frontière (BGP, *Border Gateway Protocol*) est un protocole d'acheminement inter systèmes autonomes conçu pour les internets TCP/IP. Normalement, tous les locuteurs BGP au sein d'un seul AS doivent être pleinement maillés afin que toutes les informations d'acheminement externes soient redistribuées à tous les autres routeurs au sein de ce système autonome (AS, *Autonomous System*). Cela représente un sérieux problème d'adaptabilité qui a été bien documenté par plusieurs propositions concurrentes.

Le présent document décrit l'utilisation et la conception d'une méthode appelée "réflexion de chemin" pour répondre au besoin de "maillage complet" de BGP interne (IBGP, *Internal BGP*).

Le présent document rend obsolètes la RFC 2796 et la RFC 1966.

Table des matières

1. Introduction

Normalement, tous les locuteurs BGP au sein d'un seul AS doivent être pleinement maillés et toutes les informations

d'acheminement externes doivent être redistribuées à tous les autres routeurs au sein de cet AS. Pour n locuteurs BGP au sein d'un AS cela exige de tenir n*(n-1)/2 sessions internes BGP (IBGP) uniques. Cette exigence de "plein maillage" ne s'adapte clairement pas quand il y a un grand nombre de locuteurs IBGP dont chacun échange un gros volume d'informations d'acheminement, comme c'est courant dans beaucoup de réseaux d'aujourd'hui.

Ce problème d'adaptation a été bien documenté, et un certain nombre de propositions ont été faites pour traiter cela [RFC4223], [RFC3065]. Le présent document représente une autre solution de remplacement pour contourner le besoin d'un "maillage complet" qu'on appelle "réflexion de chemin". Cette approche permet à un locuteur BGP (appelé un "réflecteur de chemin") d'annoncer les chemins IBGP appris à certains homologues IBGP. Elle représente un changement du concept couramment compris de IBGP, et l'ajout de deux nouveaux attributs BGP facultatifs non transitifs pour empêcher les boucles dans les mises à jour d'acheminement.

Le présent document rend obsolète les [RFC2796] et [RFC1966].

2. Spécification des exigences

Les mots clés "DOIT", "NE DOIT PAS", "EXIGE", "DEVRA", "NE DEVRA PAS", "DEVRAIT", "NE DEVRAIT PAS", "RECOMMANDE", "PEUT", et "FACULTATIF" en majuscules dans ce document sont à interpréter comme décrit dans le BCP 14, [RFC2119].

3. Critères de conception

La réflexion de chemin a été conçue pour satisfaire les critères suivants :
- o Simplicité : toute solution de remplacement doit être simple à configurer et facile à comprendre.
- o Facilité de transition : il doit être possible de transiter d'une configuration de maillage complet sans avoir besoin de changer la topologie ou l'AS. C'est une surcharge de gestion malencontreuse de la technique proposée dans la [RFC3065].
- o Compatibilité : il doit être possible à des homologues non conformes à IBGP de continuer de faire partie de l'AS ou domaine original sans aucune perte des informations d'acheminement de BGP.

Ces critères ont été motivés par les expériences de fonctionnement de très grands réseaux topologiquement riches avec de nombreuses connexions externes.

4. Réflexion de chemin

L'idée de base de la réflexion de chemin est très simple. Considérons l'exemple simple décrit par la Figure 1.

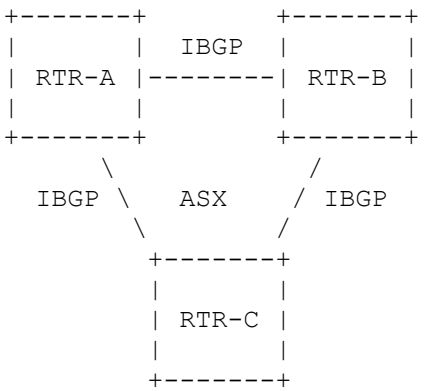

Figure 1 : Maillage complet IBGP

Dans ASX, il y a trois locuteurs IBGP (les routeurs RTR-A, RTR-B, et RTR-C). Avec le modèle BGP existant, si RTR-A reçoit un chemin externe et qu'il est choisi comme meilleur chemin, il doit annoncer le chemin externe aux deux RTR-B et RTR-C. RTR-B et RTR-C (en tant que locuteurs IBGP) ne vont pas réannoncer ces chemins appris par IBGP aux autres

locuteurs IBGP.

Si cette règle est relâchée et si il est permis au RTR-C d'annoncer les chemins appris par IBGP aux homologues IBGP, il pourrait alors réannoncer (ou refléter) les chemins IBGP appris de RTR-A à RTR-B et vice versa. Cela éliminerait le besoin d'une session IBGP entre RTR-A et RTR-B comme montré à la Figure 2.

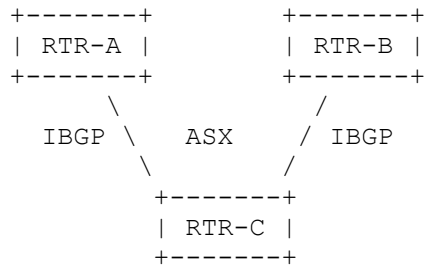

Figure 2 : Réflexion de chemin IBGP

Le schéma de réflexion de chemin se fonde sur ce principe de base.

5. Terminologie et concepts

On utilise le terme de "réflexion de chemin" pour décrire l'opération par laquelle un locuteur BGP annonce un chemin appris par IBGP à un autre homologue IBGP. Un tel locuteur BGP est dit être un "réflecteur de chemin" (RR, *Route Reflector*) et un tel chemin est appelé un chemin réfléchi.

Les homologues internes d'un RR sont divisés en deux groupes :
1) les homologues clients
2) les homologues non clients

Un RR reflète les chemins entre ces groupes, et peut refléter des chemins aux homologues clients. Un RR avec ses homologues clients forment une grappe. L'homologue non client doit être pleinement maillé mais les homologues clients n'ont pas besoin de l'être. La Figure 3 décrit un exemple simple qui souligne les composants de base de RR en utilisant la terminologie notée ci-dessus.

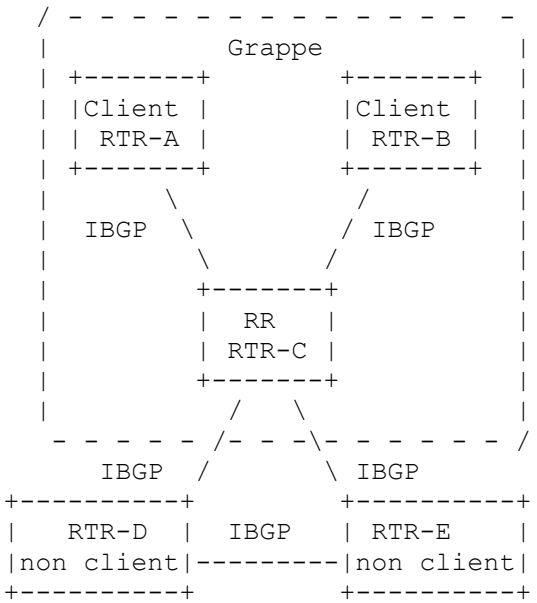

Figure 3 : Composants de RR

6. Fonctionnement

Quand un RR reçoit un chemin d'un homologue IBGP, il choisit le meilleur chemin sur la base de sa règle de sélection de chemin. Après le choix du meilleur chemin, il doit faire ce qui suit selon le type d'homologue d'où il a reçu le meilleur chemin :

1) Chemin provenant d'un homologue non client : refléter à tous les clients.
2) Chemin provenant d'un homologue client : refléter à tous les homologues non clients et aussi aux homologues clients. (Donc, les homologues clients ne sont pas obligés d'être pleinement maillés.)

Un système autonome pourrait avoir de nombreux RR. Un RR traite les autres RR juste comme tout autre locuteur BGP interne. Un RR pourrait être configuré à avoir d'autres RR dans un groupe de clients ou un groupe de non clients.

Dans une configuration simple, le sous-système réseau pourrait être divisé en de nombreuses grappes. Chaque RR serait configuré avec les autres RR comme homologues non clients (donc tous les RR vont être pleinement maillés). Les clients vont être configurés à ne maintenir la session IBGP qu'avec le RR dans leur grappe. Du fait de la réflexion de chemin, tous les locuteurs IBGP vont recevoir les informations d'acheminement réfléchies.

Il est possible dans un système autonome d'avoir des locuteurs BGP qui ne comprennent pas le concept de réflecteurs de chemin (appelons les des locuteurs BGP conventionnels). Le schéma de réflecteur de chemin permet à de tels locuteurs BGP conventionnels de coexister. Les locuteurs BGP conventionnels pourraient être membres d'un groupe de non clients ou d'un groupe ce clients. Cela permet une migration aisée et graduelle du modèle IBGP actuel au modèle de réflexion de chemin. On pourrait commencer à créer des grappes en configurant un seul routeur comme RR désigné et en configurant les autres RR et leurs clients comme des homologues IBGP normaux. Des grappes supplémentaires peuvent être créées graduellement.

7. RR redondants

Généralement, une grappe de clients va avoir un seul RR. Dans ce cas, la grappe sera identifiée par l'identifiant BGP du RR. Cependant, cela représente un seul point de défaillance, de sorte que pour rendre possible d'avoir plusieurs RR dans la même grappe, tous les RR dans la même grappe peuvent être configurés avec un CLUSTER_ID *(identifiant de grappe)* de quatre octets afin qu'un RR puisse éliminer les chemins provenant des autres RR dans la même grappe.

8. Éviter les boucles d'informations d'acheminement

Quand un chemin est réfléchi, il est possible que, suite à une mauvaise configuration, se forment des boucles de redistribution de chemins. La méthode de réflexion de chemin définit les attributs suivants pour détecter et éviter des boucles d'informations d'acheminement :

ORIGINATOR_ID *(identifiant du générateur)* : ORIGINATOR_ID est un nouvel attribut BGP facultatif, non transitif dont le code de type est 9. Cet attribut est long de quatre octets et sera créé par un RR quand il réfléchit un chemin. Cet attribut va porter l'identifiant BGP du générateur du chemin dans l'AS local. Un locuteur BGP NE DEVRAIT PAS créer un attribut ORIGINATOR_ID si il en existe déjà un. Un routeur qui reconnaît l'attribut ORIGINATOR_ID DEVRAIT ignorer un chemin reçu avec son identifiant BGP comme ORIGINATOR_ID.

CLUSTER_LIST *(liste de grappes)* : CLUSTER_LIST est un nouvel attribut BGP facultatif, non transitif dont le code de type est 10. C'est une séquence de valeurs de CLUSTER_ID représentant le chemin réfléchi parcouru.

Quand un RR reflète un chemin, il DOIT ajouter l'identifiant de grappe locale à la liste des grappes. Si la liste des grappes est vide, il DOIT en créer une nouvelle. En utilisant cet attribut, un RR peut identifier si les informations d'acheminement ont formé une boucle revenant à la même grappe à cause d'une mauvaise configuration. Si l'identifiant de grappe local se trouve dans la liste des grappes, l'annonce reçue DEVRAIT être ignorée.

9. Impact sur le choix de chemin

Les règles de départage du processus de décision de BGP (au paragraphe 9.1.2.2 de la [RFC4271]) sont modifiées comme

suit :

Si un chemin porte l'attribut ORIGINATOR_ID, alors dans l'étape f) le ORIGINATOR_ID DEVRAIT être traité comme l'identifiant BGP du locuteur BGP qui a annoncé le chemin.

De plus, la règle suivante DEVRAIT être insérée entre les étapes f) et g) : un locuteur BGP DEVRAIT préférer un chemin avec la plus courte longueur de CLUSTER_LIST. La longueur de CLUSTER_LIST est zéro si un chemin ne porte pas d'attribut CLUSTER_LIST.

10. Considérations de mise en œuvre

Il faudrait veiller à s'assurer qu'aucun des attributs de chemin BGP définis ci-dessus ne peut être modifié par configuration lors de l'échange des informations d'acheminement internes entre les RR et les clients et non clients. Leur modification pourrait résulter en boucles d'acheminement.

De plus, quand un RR reflète un chemin, il NE DEVRAIT PAS modifier les attributs de chemin NEXT_HOP, AS_PATH, LOCAL_PREF, et MED. Leur modification pourrait résulter en boucles d'acheminement.

11. Considérations de configuration et de déploiement

Le protocole BGP ne fournit aucun moyen pour qu'un client s'identifie dynamiquement comme client d'un RR. La façon la plus simple de réaliser cela est la configuration manuelle.

Un des composants clés de l'approche par la réflexion de chemin du traitement du problème de l'adaptabilité est que le RR résume les informations de chemin et ne réfléchit que le meilleur chemin.

Les deux métriques de discriminant multi-sorties (MED, *Multi-Exit Discriminator*) et de protocole de passerelle intérieure (IGP, *Interior Gateway Protocol*) peuvent impacter le choix de chemin BGP. Parce que les MED ne sont pas toujours comparables et que la métrique de IGP peut différer pour chaque routeur, avec certaines topologies de réflexion de chemin l'approche de la réflexion de chemin peut ne pas donner le même résultat de choix de chemin que l'approche du maillage IBGP complet. Une façon de rendre le choix de chemin le même que ce qu'il serait avec l'approche du maillage IBGP complet est de s'assurer que les réflecteurs de chemins ne sont jamais forcés d'effectuer le choix de chemin BGP sur la base d'une métrique IGP qui serait significativement différente de la métrique IGP de leurs clients, ou sur la base de MED incomparables. Le premier peut se faire en configurant la métrique IGP intra grappe à être meilleure que la métrique IGP inter grappes, et en maintenant un maillage complet au sein de la grappe. Le dernier peut être réalisé en :

- réglant la préférence locale d'un chemin au routeur bordure à refléter les valeurs de MED, ou
- de s'assurer que les longueurs de chemin d'AS provenant des différents AS sont différentes quand la longueur du chemin d'AS est utilisée comme critère de choix de chemin, ou
- en configurant les politiques fondées sur la communauté à influencer le choix de chemin.

On pourrait objecter cependant que la dernière exigence est trop restrictive, et peut-être impraticable dans certains cas. On pourrait aussi objecter que tant qu'il n'y a pas de boucle d'acheminement, il n'y a pas de raison impérieuse de forcer le choix de chemin avec les réflecteurs de chemin à être le même que ce qu'il serait avec l'approche du maillage IBGP complet.

Pour empêcher les boucles d'acheminement et conserver une vue cohérente de l'acheminement, il est essentiel que la topologie du réseau soit considérée avec soin lors de la conception d'une topologie de réflexion de chemin. En général, la topologie de réflexion de chemin devrait être congruente avec la topologie du réseau quand il existe plusieurs chemins pour un préfixe. Une approche couramment utilisée est la réflexion fondée sur un point de présence (POP, *Point of Presence*) dans laquelle chaque POP tient ses propres réflecteurs de chemin qui desservent les clients dans le POP, et dont tous les réflecteurs de chemin sont pleinement maillés. De plus, les clients des réflecteurs dans chaque POP sont souvent pleinement maillés pour les besoins de l'acheminement optimal intra POP, et les métriques intra POP IGP sont configurées à être meilleures que les métriques inter POP IGP.

12. Considérations sur la sécurité

Cette extension à BGP ne change pas les problèmes de sécurité sous-jacents inhérents à l'IBGP [RFC2385], [RFC4271] existant.

13. Remerciements

Les auteurs tiennent à remercier Dennis Ferguson, John Scudder, Paul Traina, et Tony Li des nombreuses discussions qui ont conduit au présent travail. Cette idée a été développée à partir d'une discussion antérieure entre Tony Li et Dimitri Haskin.

De plus, les auteurs souhaitent rendre hommage au précieux travail de relecture et aux suggestions de Yakov Rekhter sur ce document, ainsi qu'aux utiles commentaires de Tony Li, Rohit Dube, John Scudder, et Bruce Cole.

14. Références

14.1 Références normatives

[RFC2119] S. Bradner, "Mots clés à utiliser dans les RFC pour indiquer les niveaux d'exigence", BCP 14, mars 1997. *(MàJ par* RFC8174)

[RFC4271] Y. Rekhter, T. Li et S. Hares, "Protocole de routeur frontière version 4 (BGP-4)", janvier 2006. *(D.S.) (MàJ par* RFC6608, RFC8212)

14.2 Références pour information

[RFC1966] T. Bates, R. Chandra, "Acheminement BGP par réflexion : une solution de remplacement au maillage IBGP intégral", juin 1996. (*Obsolète, voir* RFC4456) (*MàJ par* RFC2796) *(Expérimentale)*

[RFC2385] A. Heffernan, "Protection des sessions de BGP via l'option de signature MD5 de TCP", août 1998. *(P.S. ; MàJ par la* RFC6691) *; remplacée par* RFC5925)

[RFC2796] T. Bates, R. Chandra, E. Chen, "Réflexion de chemin BGP - une alternative à IBGP à maillage complet", avril 2000. (*Obsolète, voir* RFC4456) *(P.S.)*

[RFC3065] P. Traina, D. McPherson, J. Scudder, "Confédérations de systèmes autonomes pour BGP", février 2001. (*Obsolète, voir* RFC5065) *(P.S.)*

[RFC4223] P. Savola, "Reclassement de la RFC 1863 comme Historique", octobre 2005. *(Information)*

Annexe A. Changements par rapport à la RFC 2796

L'impact sur le choix de chemin a été ajouté.

La description du codage de l'attribut CLUSTER_LIST a été retirée car elle est redondante avec celle de la spécification BGP, et le champ Longueur d'attribut y est par erreur décrit comme d'un octet.

Annexe B. Changements par rapport à la RFC 1966

Les changements cités ci-dessus, plus ce qui suit.

La terminologie relative au choix de chemin est précisée, et la référence aux chemins/homologues EBGP a été retirée.

Le traitement de boucle d'informations d'acheminement (due à la réflexion de chemin) par un receveur est précisé et rendu plus cohérent.

L'ajout d'un CLUSTER_ID à la CLUSTER_LIST a été changé de "ajout" en "ajout en préfixe" pour refléter le code déployé.

Le paragraphe "Considérations de configuration et de déploiement" a été étendu pour traiter plusieurs questions de fonctionnement.

Adresse des auteurs

Tony Bates
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
mél : tbates@cisco.com

Ravi Chandra
Sonoa Systems, Inc.
3255-7 Scott Blvd.
Santa Clara, CA 95054
mél : rchandra@sonoasystems.com

Enke Chen
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
mél : enkechen@cisco.com

Déclaration complète de droits de reproduction

Remerciement
Le financement de la fonction d'édition des RFC est fourni par l'activité de soutien administatif (IASA) de l'IETF.