Groupe de travail Réseau
Request for Comments : 4684
RFC mise à jour : 4364
Catégorie : Sur la voie de la normalisation
Traduction Claude Brière de L'Isle

P. Marques, Juniper Networks
R. Bonica, Juniper Networks
L. Fang, Cisco Systems, Inc.
L. Martini, Cisco Systems, Inc.
R. Raszuk, Cisco Systems, Inc.
K. Patel, Cisco Systems, Inc.
J. Guichard, Cisco Systems, Inc.
novembre 2006

Distribution de chemin contraint pour réseaux privés virtuels (VPN) au protocole Internet selon le protocole de routeur frontière/commutation d'étiquettes multi protocoles (BGP/MPLS)

Résumé
Le présent document définit les procédures de BGP multi protocoles (MP-BGP, *Multi-Protocol BGP*) qui permettent aux locuteurs BGP d'échanger des informations d'accessibilité de chemin cible. Ces informations peuvent être utilisées pour construire un graphe de distribution de chemins afin de limiter la propagation des informations d'accessibilité de couche réseau (NLRI, *Network Layer Reachability Information*) de réseau privé virtuel (VPN, *Virtual Private Network*) entre différents systèmes autonomes ou grappes distinctes du même système autonome. Le présent document met à jour la RFC4364.

Table des matières

1. Introduction

Dans les VPN IP BGP/MPLS, les routeurs PE utilisent les communautés étendues de chemin cible (RT, *Route Target*) pour contrôler la distribution des routes dans les VRF. Au sein d'un certain maillage iBGP, les routeurs PE ont seulement besoin de détenir les routes marquées avec des cibles de chemin relevant des VRF qui ont des rattachements CE locaux.

Il est cependant courant pour un système autonome d'utiliser la réflexion de chemin [RFC4456] afin de simplifier le processus d'activation d'un nouveau routeur PE dans le réseau et de limiter la taille du maillage d'échangeurs de trafic iBGP.

Dans un tel scénario, ainsi que quand des VPN peuvent avoir des membres dans plus d'un système autonome, le nombre de routes portées par les routeurs inter grappes ou inter AS est une considération importante.

Afin de limiter les informations d'acheminement de VPN qui sont conservées dans un certain réflecteur de chemins, la [RFC4364] suggère au paragraphe 4.3.3 l'utilisation d'un "filtrage de chemins coopératif" [RFC5291] entre les réflecteurs de chemins. Le présent document étend le travail de filtrage de chemin sortant (ORF, *Outbound Route Filtering*) [RFC4364] pour inclure la prise en charge de plusieurs systèmes autonomes et topologies de VPN asymétriques telles que noyaux et lignes de collecte et de distribution.

Bien qu'il serait possible d'étendre le codage actuellement défini pour l'ORF de communauté étendue afin de le faire, BGP a déjà lui-même toute la machinerie nécessaire pour disséminer des informations arbitraire sans dépendre de la boucle, à la fois dans un seul système autonome et à travers plusieurs systèmes autonomes.

Le présent document s'appuie sur le modèle décrit dans la [RFC4364] et sur le concept de filtrage coopératif de chemin en ajoutant la capacité de propager les informations de membre de chemin cible entre les maillages iBGP. Il est conçu pour remplacer le "filtrage coopératif de chemin" pour les applications en rapport avec les VPN.

En utilisant les messages MP-BGP UPDATE pour propager les informations de membre de chemin cible, il est possible de réutiliser toute cette machinerie, incluant la réflexion de chemin, les confédérations, et la détection de boucle d'informations inter AS.

Les informations de membre de chemin cible reçues peuvent alors être utilisées pour restreindre les annonces de NLRI de VPN aux homologues qui ont annoncé leurs chemins cibles respectifs, construisant effectivement un graphe de distribution de chemins. Dans ce modèle, les informations d'acheminement de NLRI de VPN s'écoulent dans la direction inverse des informations de membre de chemin cible.

Ce mécanisme est applicable à toutes les NLRI BGP qui contrôlent la distribution des informations d'acheminement en utilisant les chemins cibles, comme VPLS [RFC4761].

Dans le présent document, le terme de NLRI (*Network Layer Reachability Information*, informations d'accessibilité de couche réseau) est utilisé pour décrire les informations d'acheminement portées via les mises à jour MP-BGP sans aucune hypothèse de sémantique.

Des NLRI consistant en {n° d'AS d'origine, chemin cible} vont être appelées des informations de membre RT pour les besoins des explications de ce document.

1.1 Terminologie

Le présent document utilise un certain nombre de termes et acronymes spécifiques des VPN provisionnés par un fournisseur, incluant ceux spécifiques de L2VPN, L3VPN et BGP. Les définitions de beaucoup de ces termes peuvent être trouvées dans le document de terminologie de VPN [RFC4026]. Cette section inclut aussi des expansions d'acronymes et de terminologie pour aider le lecteur.

AFI *(Address Family Identifier)* : identifiant de famille d'adresse (type d'adresse BGP)
BGP (Border Gateway Protocol) : protocole de passerelle frontière
CE *(Customer Edge)* = extrémité client (routeur)
iBGP *(Internal BGP)* : BGP interne (c'est-à-dire, une session d'échange de trafic BGP qui connecte deux routeurs au sein d'un système autonome)
L2VPN *(Layer 2 Virtual Private Network)* : réseau privé virtuel de couche 2
L3VPN (*Layer 3 Virtual Private Network)* :réseau privé virtuel de couche 3
MP-BGP *(MultiProtocol-Border Gateway Protocol)* : protocole de passerelle frontière multi protocoles
MPLS *(MultiProtocol Label Switching)* : commutation d'étiquettes multi protocoles
NLRI *(Network Layer Reachability Information)* : informations d'accessibilité de couche réseau
ORF *(Outbound Route Filtering)* : filtrage de chemin sortant
PE *(Provider Edge)* : côté fournisseur (routeur)

RT (Route Target) : chemin cible (c'est-à-dire, communauté BGP étendue qui conditionne les informations d'accessibilité de couche réseau à la qualité de VPN membre)
SAFI *(Subsequent Address Family Identifier)* = identifiant de la famille d'adresse suivante (sous type d'adresse BGP)
VPLS *(Virtual Private LAN Service)* : service de LAN virtuel privé
VPN *(Virtual Private Network)* : réseau privé virtuel
VPN BGP/MPLS : mise en œuvre de VPN de couche 3 fondée sur BGP et MPLS

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. Distribution de NLRI

3.1 Distribution de chemin de VPN inter AS

Afin de mieux comprendre le problème posé, il sera utile de le diviser en ses composants inter système autonome (AS) et intra AS. La Figure 1 représente un graphe arbitraire de systèmes autonomes (de a à j) interconnectés de façon ad hoc. La discussion qui suit ignore la complexité de la distribution de chemin intra AS.

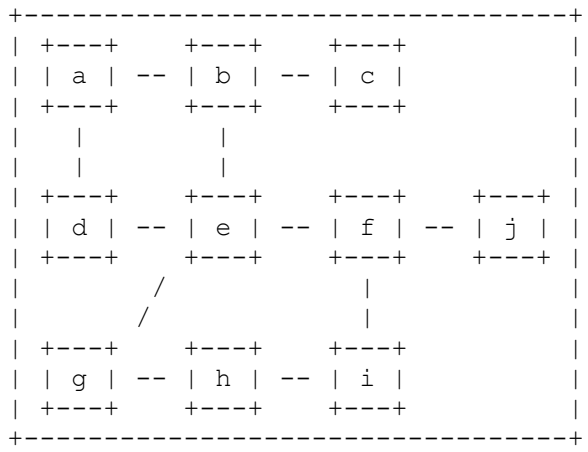


Figure 1 : Topologie de systèmes autonomes

Considérons le cas simple d'un VPN avec des rattachements CE dans les AS a et i qui utilise une seule route cible pour contrôler la distribution de chemins de VPN. Idéalement, on aimerait construire un graphe d'écoulement pour les chemins de VPN respectifs qui n'incluraient pas les nœuds (c, g, h, j). Les nœuds (c, j) sont des AS d'extrémité qui n'exigent pas ces informations, tandis que les nœuds (g, h) ne sont pas dans le plus court chemin inter AS entre (e) et (i) et donc devraient être exclus via le choix standard de chemin BGP.

Afin de réaliser cela, on va s'appuyer sur les AS a et AS i, générant une NLRI consistant en {n° d'AS d'origine, chemin cible} (informations de membre RT). La réception d'une telle annonce par un des AS dans le réseau va signaler le besoin de distribuer des chemins de VPN contenant cette communauté de chemins cibles à l'homologue qui a annoncé cette route.

En utilisant les informations de membre RT qui incluent à la fois le chemin cible et le numéro d'AS d'origine on permet aux locuteurs BGP d'utiliser les règles standard de choix de chemin concernant la longueur du chemin d'AS (et les autres mécanismes de politique) pour élaguer les chemins dupliqués dans le graphe d'arrosage des informations de membre RT, tout en conservant les informations requises pour atteindre tous les systèmes autonomes qui annoncent la chemin cible.

Dans l'exemple ci-dessus, l'AS e a besoin de conserver un chemin pour l'AS a afin d'arroser les informations d'acheminement de VPN originaires de AS i et vice-versa. Il devrait cependant, comme politique par défaut, élaguer les chemins moins préférés comme le plus long chemin pour AS i avec le chemin d'AS (g h i).

En étendant l'exemple ci-dessus pour inclure l'AS j comme membre du graphe de distribution de VPN on causerait l'annonce par l'AS f de deux NLRI de membre RT à l'AS e, une contenant l'origine AS i et une contenant l'origine AS j. Bien qu'annoncer un seul chemin serait suffisant pour garantir que les informations de VPN parviennent à tous les AS membres du VPN, ce n'est pas assez pour les choix de chemin désirés. Dans l'exemple ci-dessus, on suppose que (f j) est choisi et annoncé. Si c'est le cas, les informations concernant le chemin (f i), qui sont nécessaires pour élaguer l'arc (e g h i) du graphe de distribution de chemins, vont manquer.

Comme avec d'autres approches pour construire les graphes de distribution, les avantages de ce mécanisme sont directement proportionnels à la "dispersion" des membres du VPN. Le comportement standard inter AS de la [RFC2547] peut être vu comme une approche de mode dense, pour faire une analogie avec les protocoles d'acheminement de diffusion groupée.

3.2 Distribution de chemin de VPN intra AS

Comme indiqué ci-dessus, le graphe de distribution de chemin de VPN inter AS, pour un certain chemin cible, est construit en créant un arc dirigé dans la direction inverse des UPDATE de membre de chemin cible reçus qui contiennent une NLRI de la forme {n° d'AS d'origine, chemin cible}.

Dans la topologie BGP d'un certain système autonome, en ce qui concerne les informations de membre RT externe (chemins cibles où le n° d'AS n'est pas l'AS local) il est aisé de voir que les règles standard de choix et d'annonce de chemin BGP [RFC4271] vont permettre à un AS de transit de créer l'état d'arrosage nécessaire.

Considérons un préfixe de NLRI IPv4, dont la source est un seul AS, qui est distribué via BGP dans un certain AS de transit. Les règles du protocole BGP garantissent qu'un locuteur BGP a un chemin valide qui peut être utilisé pour transmettre des paquets de données pour ce préfixe de destination, dans le chemin inverse des mises à jour d'acheminement reçues.

Par le même jeton, et étant donné qu'une clé {n° d'AS d'origine, chemin cible} assure l'unicité entre plusieurs AS qui peuvent être à la source de ce chemin cible, les procédures de choix et d'annonce de chemin BGP garantissent qu'un chemin valide de distribution de chemin de VPN existe pour l'origine de l'annonce des informations de membre du chemin cible.

Les informations de membre du chemin cible qui ont leur origine dans le système autonome exigent cependant un examen plus attentif. Plusieurs routeurs PE au sein d'un certain système autonome peuvent générer les mêmes NLRI {n° d'AS d'origine, chemin cible}, et donc les règles par défaut d'annonce de chemin ne sont plus suffisantes pour garantir qu'au sein de cet AS chaque nœud dans le graphe de distribution a choisi un chemin faisable pour chacun des PE qui importent le chemin cible donné.

Quand on traite les NLRI de membre RT reçues d'homologues iBGP internes, il est nécessaire de considérer tous les chemins iBGP disponibles pour un certain préfixe de RT, pour construire le filtre de chemin sortant, et pas juste le meilleur chemin.

De plus, quand il annonce les informations de membre de chemin cible générées par le système autonome local à un homologue iBGP, un locuteur BGP devra modifier sa procédure pour calculer les attributs BGP de façon que ce qui suit s'applique :

i. Quand il annonce les NLRI de membre RT à un client réflecteur de chemin, l'attribut Origine devra être établi à l'identifiant de routeur de l'annonceur, et l'attribut Prochain bond devra être réglé à l'adresse locale pour cette session.

ii. Quand il annonce les NLRI de membre RT à un homologue non client, si le meilleur chemin choisi par la procédure de choix de chemin décrite au paragraphe 9.1 de la spécification BGP de base [RFC4271] est une route reçue d'un homologue non client, et si il y a un autre chemin pour la même destination venant d'un client, les attributs du chemin de client sont annoncés à l'homologue.

La première de ces règles d'annonce de chemin est conçue de façon telle que l'origine des NLRI de membre de RT n'élimine pas une NLRI de membre de RT qui lui est réfléchie, permettant donc au réflecteur de chemin d'utiliser cette NLRI de membre de RT afin de signaler au client qu'il devrait distribuer les chemins de VPN avec la cible spécifique vers le réflecteur.

La second règle permet à tout locuteur BGP présent dans un maillage iBGP de signaler l'intérêt de ses clients de réflexion de chemin à recevoir des chemins de VPN pour cette cible.

Ces procédures supposent que la topologie de réflexion de chemin du système autonome est configurée de façon telle que l'acheminement d'envoi individuel IPv4 fonctionne correctement. Par exemple, les grappes de réflexion de chemin doivent être contiguës.

Une solution de remplacement à la procédure donnée ci-dessus serait d'avoir des sources de chemin différentes par PE, comme des NLRI de la forme {identifiant d'origine, chemin cible}, et de les agréger à la bordure du réseau. La solution adoptée est considérée comme plus avantageuse que cette dernière en ce qu'elle exige moins d'informations d'acheminement au sein d'un AS donné.

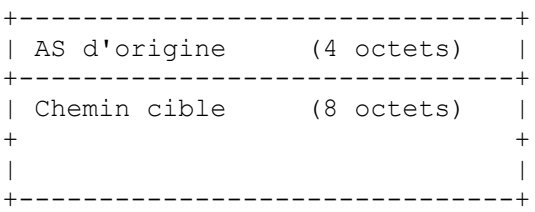
4. Annonces de NLRI de membre de chemin cible

Les NLRI de membre de RT sont annoncées dans les messages BGP UPDATE en utilisant les attributs MP_REACH_NLRI et MP_UNREACH_NLRI [RFC2858]. La paire de valeurs [AFI, SAFI] utilisée pour identifier ces NLRI est (AFI=1, SAFI=132).

Le champ Prochain bond de l'attribut MP_REACH_NLRI devra être interprété comme une adresse IPv4 chaque fois que la longueur de l'adresse de prochain bond est de 4 octets, et comme une adresse IPv6 chaque fois que la longueur de l'adresse de prochain bond fait 16 octets.

Le champ NLRI dans les MP_REACH_NLRI et MP_UNREACH_NLRI est un préfixe de 0 à 96 bits, codé comme défini à la Section 4 de la [RFC2858].

Ce préfixe est structuré comme suit :



Sauf pour le chemin cible par défaut, qui est codé comme un préfixe de longueur zéro, la longueur minimum de préfixe est de 32 bits, car le champ AS d'origine ne peut pas être interprété comme un préfixe.

Les chemins cibles peuvent alors être exprimés comme des préfixes, où, par exemple, un préfixe va englober toutes les communautés étendues de chemin cible allouées par un certain administrateur global [RFC4360].

Le chemin cible par défaut peut être utilisé pour indiquer à un homologue la volonté de recevoir toutes les annonces de chemin de VPN comme, par exemple, le cas d'un réflecteur de chemin qui parle à un de ses clients de routeur PE.

5. Annonce de capacité

Un locuteur BGP qui souhaite échanger des informations de membre RT doit utiliser le code de capacité d'extension multi protocoles, comme défini dans la [RFC2858], pour annoncer la paire (AFI, SAFI) correspondante.

Un locuteur BGP PEUT participer à la distribution des informations de RT sans utiliser les informations apprises pour les besoins du filtrage de chemin de sortie de NLRI de VPN, bien que ce soit déconseillé.

6. Fonctionnement

Un chemin de NLRI de VPN devrait être annoncé à un homologue qui participe à l'échange d'informations de membre de route cible si cet homologie a annoncé soit les NLRI de membre de route cible par défaut soit des NLRI de membre de route cible contenant une des cibles contenues dans l'attribut communautés étendues du chemin de VPN en question.

Quand un locuteur BGP reçoit un UPDATE BGP qui annonce ou retire une certaine NLRI de membre de route cible, il devrait examiner les RIB-OUT des NLRI de VPN et réévaluer l'état d'annonce des chemins qui correspondent à la route cible en question.

Un locuteur BGP devrait générer l'ensemble minimum de mises à jour de chemins de VPN BGP (annonces et/ou retraits) nécessaires pour la transition entre l'état antérieur et l'état en cours du graphe de distribution de chemins qui est déduit des informations de membre de route cible.

Comme indication que l'échange initial de membres de RT est achevé, les mises en œuvre DEVRAIENT générer un marqueur de fin de RIB, comme défini dans la [RFC4724], pour le (afi, safi) de membre de route cible, sans considération de si le redémarrage en douceur est activé sur la session BGP. Cela permet au receveur de savoir quand il a reçu tout le contenu des informations de membre de l'homologue. L'échange de NLRI de VPN devrait suivre la réception des marqueurs de fin de RIB.

Si un locuteur BGP choisit de retarder l'annonce des mises à jour de chemins de VPN BGP jusqu'à ce qu'il reçoive le marqueur de fin de RIB, il DOIT établir une limite supérieure de ce délai. Par défaut, une valeur de 60 secondes devrait être utilisée.

7. Considérations de déploiement

Ce mécanisme réduit les exigences d'adaptabilité qui sont imposées aux réflecteurs de chemins en limitant le nombre de chemins et d'événements de VPN qu'un réflecteur doit traiter aux chemins de VPN utilisés par ses clients directs. Par défaut, un réflecteur doit s'adapter au nombre total de chemins de VPN présents sur le réseau.

Cela signifie aussi qu'il est maintenant possible de réduire la charge imposée à un certain réflecteur en divisant les routeurs PE présents sur sa grappe dans un nouvel ensemble de grappes. C'est un changement de configuration localisé qui ne doit pas affecter de système en dehors de cette grappe.

L'efficacité du filtrage fondé sur la route cible dépend de la dispersion des membres du VPN.

Les mêmes mécanismes de politique applicables aux autres NLRI sont aussi applicables aux informations de membre de RT. Cela donne à un opérateur de réseau l'option de contrôler quelles routes de VPN sont annoncées dans une bordure inter domaines en filtrant les annonces entrantes de membre RT acceptables.

Par exemple, dans le cas inter AS, il est probable qu'un certain VPN est seulement connecté à un sous ensemble de tous les AS participants. Le seul mécanisme actuel pour limiter la portée de l'arrosage de chemins de VPN est le filtrage manuel sur les routeurs de bordure externe BGP. Avec la proposition actuelle, un tel filtrage peut être effectué en accord avec les informations dynamiques de membre de route cible.

Dans certains déploiements inter AS, tous les RT utilisés pour un certain VPN n'ont pas de signification externe. Par exemple, un VPN peut utiliser un noyau RT et un rayon RT en interne à un système autonome. Le rayon RT n'a pas de signification en dehors de cet AS, de sorte qu'il peut être supprimé à un routeur bordure externe. Les mêmes règles de politique qui résultent en un filtrage de communauté étendue peuvent être appliquées aux informations de membre de RT afin d'éviter d'annoncer des NLRI de membre de RT pour le rayon RT dans l'exemple ci-dessus.

Dans le présent document, on suppose que les systèmes autonomes s'accordent sur une convention d'allocation de RT. La traduction de RT à la frontière du routeur bordure externe est considérée comme une décision de mise en œuvre locale, car elle ne devrait pas affecter l'inter opérabilité.

8. Considérations sur la sécurité

Le présent document n'altère pas les propriétés de sécurité des VPN fondés sur BGP. Cependant, on notera que les filtres de chemin de sortie construits à partir des NLRI d'informations de membre RT ne sont pas destinés à des buts de sécurité. Lors de l'échange des informations d'acheminement entre des domaines administratifs séparés, il est de bonne pratique de filtrer toutes les NLRI entrantes et sortantes par d'autres moyens en plus des informations de membre RT. Les mises en œuvre DEVRAIENT aussi fournir des moyens de filtrer les informations de membre RT.

9. Remerciements

La présente proposition s'appuie sur le mécanisme de filtrage de chemin de communauté étendue défini dans la [RFC5291]. Ahmed Guetari a joué un rôle décisif dans la définition des exigences pour cette proposition. Les auteurs tiennent aussi à remercier Yakov Rekhter, Dan Tappan, Dave Ward, John Scudder, et Jerry Ash de leurs commentaires et suggestions.

10. Références

10.1 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)

[RFC2858] T. Bates et autres, "Extensions multiprotocoles pour BGP-4", juin 2000. (*Obsolète, voir* RFC4760) *(P.S.)*

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

[RFC4360] S. Sangli et autres, "Attribut BGP-4 Communauté étendue", février 2006. *(P.S.)*

[RFC4364] E. Rosen et Y. Rekhter, "Réseaux privés virtuels IP BGP/MPLS", février 2006. (*P.S., MàJ par* RFC4577, RFC4684)

[RFC4456] T. Bates, E. Chen, R. Chandra, "Réflexion de chemin BGP : une solution de remplacement au BGP interne à maillage complet (IBGP)", avril 2006. (*Remplace* RFC2796, RFC1966) *(D.S.)*

10.2 Références pour information

[RFC4026] L. Andersson et T. Madsen, "Terminologie des réseaux privés virtuels (VPN) approvisionnés par le fournisseur", mars 2005.

[RFC5291] E. Chen, Y. Rekhter, "Capacité de filtrage de chemin sortant pour BGP-4", août 20080. *(P.S.)*

[RFC4724] S. Sangli et autres, "Mécanisme de redémarrage en douceur pour BGP", janvier 2007. *(P.S.)*

[RFC4761] K. Ko.mpella et Y. Rekhter, éditeurs "Service de LAN privé virtuel (VPLS) utilisant BGP pour l'auto découverte et la signalisation", janvier 2007. *(P.S. ; MàJ par* RFC8395)

Adresse des auteurs

Pedro Marques
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US
mél : roque@juniper.net

Ronald Bonica
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US
mél : rbonica@juniper.net

Jim Guichard
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough, MA 01719
US
mél : jguichar@cisco.com

Luyuan Fang
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough, MA 01719
US
mél : lufang@cisco.com

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO 80112
US
mél : lmartini@cisco.com

Robert Raszuk
Cisco Systems, Inc.
170 West Tasman Dr
San Jose, CA 95134
US
mél : rraszuk@cisco.com

Keyur Patel
Cisco Systems, Inc.
170 West Tasman Dr
San Jose, CA 95134
US
mél : keyupate@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 administratif (IASA) de l'IETF.