Groupe de travail Réseau
Request for Comments : 1772
RFC rendue obsolète : 1655
Catégorie : En cours de normalisation
Traduction Claude Brière de L'Isle

Y. Rekhter, IBM Corp.
P. Gross, MCI
éditeurs
mars 1995

Application du protocole de routeur frontière dans l'Internet

Statut de ce mémoire
Le présent document spécifie un protocole Internet en cours de normalisation pour la communauté de l'Internet, et appelle à des discussions et suggestions pour son amélioration. Prière de se référer à l'édition en cours des "Normes officielles des protocoles de l'Internet" (STD 1) pour l'état de normalisation et le statut de ce protocole. La distribution du présent mémoire n'est soumise à aucune restriction.

Résumé
Le présent document, conjointement avec le document qui l'accompagne, "Protocole 4 de routeur frontière (BGP-4)", définit un protocole d'acheminement de systèmes inter autonomes pour l'Internet. "Protocole 4 de routeur frontière (BGP-4)" définit la spécification du protocole BGP, et le présent document décrit l'usage de BGP dans l'Internet.

Des informations sur les progrès de BGP peuvent être suivies et/ou rapportées sur la liste de diffusion de BGP à (bgp@ans.net).

Table des matières

1. Introduction

Le présent mémoire décrit l'utilisation du protocole de routeur frontière (BGP, *Border Gateway Protocol*) [RFC1771] dans l'environnement de l'Internet. BGP est un protocole d'acheminement inter systèmes autonomes (SA). Les informations d'accessibilité de réseau échangées via BGP fournissent des informations suffisantes pour détecter les boucles

d'acheminement et mettre en application les décisions d'acheminement sur la base des préférences de performances et des contraintes de politiques comme indiqué dans la RFC 1104 [RFC1104]. En particulier, BGP échange des informations d'acheminement qui contiennent les chemins de SA complets et met en application les politiques d'acheminement sur la base des informations de configuration.

Comme l'Internet a évolué et s'est étendu ces dernières années, il est devenu malheureusement évident qu'il est temps de faire face à de sérieux problèmes d'adaptation. Parmi eux :

- L'épuisement de l'espace d'adresses réseau de classe B. Une cause fondamentale de ce problème est l'absence d'une classe de réseau d'une taille qui soit appropriée pour les organisations de taille moyenne ; la classe C, avec un maximum de 254 adresses d'hôtes, est trop petite, alors que la classe B, qui permet jusqu'à 65 534 adresses, est trop grande pour être remplie.
- La croissance des tableaux d'acheminement dans les routeurs Internet dépasse la capacité de gestion efficace des logiciels actuels (et des gens)).
- L'épuisement éventuel de l'espace d'adresses IP à 32 bits.

Il est devenu clair que les deux premiers de ces problèmes vont vraisemblablement devenir critiques dans les deux ou trois prochaines années. L'acheminement inter domaines sans classe (CIDR, *Classless inter-domain routing*) tente de régler ces problèmes en proposant un mécanisme pour ralentir la croissance du tableau d'acheminement et le besoin d'allouer de nouveaux numéros de réseau IP. Il n'essaye pas de résoudre le troisième problème, qui est par nature à plus long terme, mais entreprend plus d'assouplir suffisamment les difficultés à court et moyen terme en permettant à l'Internet de continuer à fonctionner efficacement pendant que des progrès sont faits pour une solution à plus long terme.

BGP-4 est une extension de BGP-3 qui fournit la prise en charge de l'agrégation et de la réduction des informations d'acheminement sur la base de l'architecture d'acheminement inter domaine sans classe (CIDR) [RFC1519]. Le présent mémoire décrit l'usage de BGP-4 dans l'Internet.

Tout l'exposé de cet article se fonde sur l'hypothèse que l'Internet est une collection de systèmes autonomes (SA) connectés de façon arbitraire. C'est-à-dire que l'Internet sera modélisé comme un graphe général dont les nœuds sont les SA et dont les bords sont les connexions entre les paires de SA.

La définition classique d'un système autonome est un ensemble de routeurs sous une seule administration technique, utilisant un protocole de passerelle intérieure et une métrique commune pour acheminer les paquets au sein du SA et utilisant un protocole de passerelle extérieure pour acheminer les paquets vers les autres SA. Depuis le développement de cette définition classique, il est devenu courant pour un seul SA d'utiliser plusieurs protocoles de passerelle intérieure et parfois plusieurs ensembles de métriques au sein d'un SA. L'utilisation du terme "système autonome" souligne le fait que, même lorsque plusieurs IGP et métriques sont utilisés, l'administration d'un SA apparaît aux autres SA comme ayant un seul plan cohérent d'acheminement intérieur et présente un tableau cohérent des destinations qui sont accessibles à travers lui.

Les SA sont supposés être administrés par une seule entité administrative, au moins pour les besoins de représentation des informations d'acheminement aux systèmes en dehors du SA.

2. Modèle topologique BGP

Lorsque nous disons qu'une connexion existe entre deux SA, nous signifions deux choses :

Une connexion physique : il y a un sous-réseau partagé de liaisons de données entre les deux SA, et sur ce sous réseau partagé, chaque SA a au moins un routeur frontière qui appartient à ce SA. Donc, le routeur frontière de chaque SA peut transmettre des paquets au routeur frontière de l'autre SA sans avoir recours à l'acheminement inter SA ou intra SA.

Une connexion BGP : il y a une session BGP entre ceux qui parlent BGP dans chacun des SA, et cette session communique les routes qui peuvent être utilisées pour des destinations spécifiques via le SA qui l'annonce.

Tout au long de ce document, nous mettons une restriction supplémentaire sur les locuteurs BGP qui forment la connexion BGP : ils doivent eux-mêmes partager le même sous-réseau de liaisons de données que partagent leurs routeurs frontières. Donc, une session BGP entre des SA adjacents n'exige aucune prise en charge de la part de l'acheminement inter SA ou intra SA. Les cas qui ne se conforment pas à cette restriction sortent du domaine d'application du présent document.

Donc, à chaque connexion, chaque SA a un ou plusieurs locuteurs BGP et un ou plusieurs routeurs frontières, et ces

locuteurs BGP et ces routeurs frontières sont tous situés sur un sous-réseau de liaisons de données partagé. Noter que les locuteurs BGP n'ont pas besoin d'être des routeurs frontières, et vice versa. Les chemins annoncés par un locuteur BGP d'un SA sur une connexion donnée sont pris comme étant faisables pour chaque routeur frontière de l'autre SA sur le même sous-réseau partagé, c'est à dire que les voisins indirects sont admis.

Beaucoup du trafic porté au sein d'un SA provient de ce SA ou s'y termine (c'est à dire que soit l'adresse IP de source, soit l'adresse IP de destination du paquet IP identifie un hôte interne à ce SA). Le trafic qui répond à cette description est appelé du "trafic local". Le trafic qui ne répond pas à cette description est appelé du "trafic de transit". Un objectif majeur de l'utilisation de BGP est de contrôler le flux de trafic de transit.

Sur la base de la façon dont un SA particulier traite le trafic de transit, le SA peut maintenant être classé dans l'une des catégories suivantes :

SA de bout : un SA qui a une seule connexion avec un autre SA. Naturellement, un SA de bout ne porte que du trafic local.

SA multi rattachements : un SA qui a des connexions avec plus d'un autre SA, mais refuse de porter du trafic de transit.

SA de transit : un SA qui a des connexions avec plus d'un autre SA, et est conçu (sous certaines restrictions de politique) pour porter à la fois du trafic de transit et du trafic local.

Comme un chemin de SA complet fournit un moyen efficace et direct pour supprimer les boucles d'acheminement et éliminer le problème du "compte à l'infini" associé à certains algorithmes de vecteur de distance, BGP n'impose pas de restrictions topologiques à l'interconnexion des SA.

3. BGP dans l'Internet

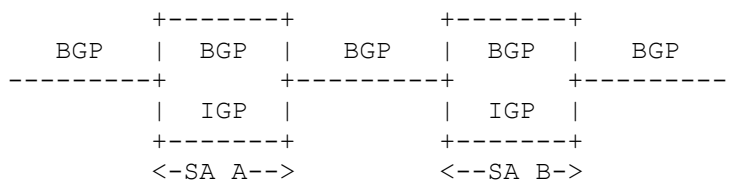
3.1 Considérations de topologie

La topologie globale de l'Internet peut être vue comme une interconnexion arbitraire de SA de transit, de multi rattachement et de bout. Afin de minimiser l'impact de l'infrastructure actuelle de l'Internet, les SA de bout et de multi rattachements n'ont pas besoin d'utiliser BGP. Ces SA peuvent fonctionner avec d'autres protocoles (par exemple, EGP) pour échanger les informations d'accessibilité avec les SA de transit. Les SA de transit qui utilisent BGP vont étiqueter ces informations comme ayant été apprises par une méthode autre que BGP. Le fait que BGP n'ait pas besoin de fonctionner sur les SA de bout ou de multi rattachement n'a pas d'impact négatif sur la qualité globale de l'acheminement inter SA pour le trafic qui est destiné ou qui vient des SA de bout ou de multi rattachements en question.

Cependant, il est recommandé que BGP soit aussi utilisé pour les SA de bout ou de multi rattachements. Dans ces situations, BGP va procurer un avantage en bande passante et en performances sur certains des protocole utilisés actuellement (tels que EGP). De plus, cela va réduire le besoin d'utiliser des chemins par défaut et va donner de meilleurs choix de chemins inter SA pour les SA à multi rattachements.

3.2 Nature globale de BGP

Au niveau global, BGP est utilisé pour distribuer les informations d'acheminement entre plusieurs systèmes autonomes. Les flux d'informations peuvent être représentés comme suit :



Ce diagramme montre que, alors que BGP porte seul les informations entre les SA, BGP et un IGP peuvent porter des informations à travers un SA. Assurer la cohérence des informations d'acheminement entre BGP et un IGP au sein d'un SA est un problème significatif qui est exposé en détails dans l'Appendice A.

3.3 Relations de voisins BGP

L'Internet est vu comme un ensemble de SA arbitrairement connectés. Les routeurs qui communiquent directement les uns avec les autres via BGP sont connus comme locuteurs BGP. Les locuteurs BGP peuvent être localisés au sein du même SA ou dans des SA différents. Les locuteurs BGP dans chaque SA communiquent les uns avec les autres pour échanger les informations d'accessibilité de réseau sur la base d'un ensemble de politiques établies au sein de chaque SA. Pour un locuteur BGP donné, un autre locuteur BGP avec lequel communique le locuteur donné est appelé un homologue externe si l'autre locuteur est dans un SA différent, alors que si l'autre locuteur est dans le même SA, il est appelé un homologue interne.

Il peut y avoir autant de locuteurs BGP qu'il paraît nécessaire au sein d'un SA. Normalement, si un SA a plusieurs connexions avec d'autres SA, plusieurs locuteurs BGP sont nécessaires. Tous les locuteurs BGP qui représentent le même SA doivent donner une image cohérente du SA à l'extérieur. Cela exige que les locuteurs BGP aient des informations d'acheminement cohérentes entre elles. Ces routeurs peuvent communiquer les uns avec les autres via BGP ou par d'autres moyens. Les contraintes de politique appliquées à tous les locuteurs BGP au sein d'un SA doivent être cohérentes. Des techniques telles que l'utilisation d'un IGP étiqueté (voir en A.2.2) peuvent être employées pour détecter de possibles incohérences.

Dans le cas d'homologues externes, l'homologue doit appartenir à un SA différent, mais partager un sous-réseau de liaisons de données commun. Ce sous réseau commun devrait être utilisé pour porter les messages BGP entre eux. L'utilisation de BGP à travers un SA interposé invalide les informations de chemin de SA. Un numéro de système autonome doit être utilisé avec BGP pour spécifier à quel système autonome appartient le locuteur BGP.

4. Exigences pour l'agrégation de chemin

Il est exigé d'une mise en œuvre conforme à BGP-4 qu'elle ait la capacité de spécifier quand un chemin agrégé peut être généré à partir d'informations d'acheminement partielles. Par exemple, un locuteur BGP à la frontière d'un système autonome (ou groupe de systèmes autonomes) doit être capable de générer un chemin agrégé pour tout un ensemble d'adresses IP de destination (dans la terminologie de BGP-4 un tel ensemble est appelé "informations d'accessibilité de la couche réseau" (NLRI, *Network Layer Reachability Information*)) sur lequel il a le contrôle administratif (y compris sur les adresses qu'il a allouées) même quand toutes ne sont pas accessibles en même temps.

Une mise en œuvre conforme peut fournir la capacité à spécifier quand des NLRI agrégées peuvent être générées.

Il est exigé d'une mise en œuvre conforme qu'elle ait la capacité de spécifier comment les NLRI peuvent être agrégées.

Il est exigé d'une mise en œuvre conforme qu'elle prenne en charge les options suivantes lorsqu'elle traite des chemins qui se chevauchent :
- installer aussi bien le chemin le moins spécifique que le plus spécifique,
- installer seulement le chemin le plus spécifique,
- installer seulement le chemin le moins spécifique,
- n'installer aucun des chemins.

Certaines politiques d'acheminement peuvent dépendre des NLRI (par exemple, "recherche" contre "commercial"). Donc, un locuteur BGP qui effectue une agrégation de chemin devrait, si possible, avoir connaissance des implications potentielles de l'agrégation des NLRI sur les politiques d'acheminement.

5. Mise en place de politiques avec BGP

BGP apporte la capacité de mettre en application des politiques sur la base de diverses préférences d'acheminement et de contraintes. Les politiques ne sont pas directement codées dans le protocole. Elles sont plutôt fournies à BGP sous la forme d'informations de configuration.

BGP met en application les politiques en jouant sur la sélection de chemins à partir d'alternatives multiples et en contrôlant la redistribution des informations d'acheminement. Les politiques sont déterminées par l'administration du SA.

Les politiques d'acheminement se rapportent à des considérations de politique, de sécurité, ou d'économie. Par exemple, si

un SA ne veut pas porter du trafic à un autre SA, il peut mettre en application une politique qui l'interdise. Les exemples suivants montrent des politiques d'acheminement qui peuvent être appliquées avec BGP :

1. Un SA multi rattachements peut refuser d'agir comme SA de transit pour d'autres SA. (Il le fait en annonçant seulement des chemins pour des destinations internes au SA.)

2. Un SA multi rattachements peut devenir un SA de transit pour un ensemble restreint de SA adjacents, c'est à dire que certains des SA, mais pas tous, peuvent utiliser le SA multi rattachements comme SA de transit. (Il le fait en annonçant ses informations d'acheminement sur cet ensemble de SA.)

3. Un SA peut favoriser ou défavoriser l'usage de certains SA pour le transport du trafic de transit provenant de lui-même.

Un certain nombre de critères en rapport avec les performances peuvent être contrôlés avec l'utilisation de BGP :

1. Un SA peut minimiser le nombre de SA de transit. (Des chemins de SA plus courts peuvent être préférés à de plus longs.)

2. La qualité du SA de transit. Si un SA détermine que deux chemins de SA, ou plus, peuvent être utilisés pour atteindre une destination donnée, ce SA peut utiliser divers moyens pour décider quel candidat chemin de SA il va utiliser. La qualité d'un SA peut être mesurée par des choses comme un diamètre, la vitesse de la liaison, la capacité, la tendance à être encombré, et la qualité du fonctionnement. Les informations sur ces qualités peuvent être déterminées par des moyens autres que BGP.

3. Préférence des chemins internes sur les chemins externes.

Par souci de cohérence au sein d'un SA, des chemins de coût égal, résultant de la combinaisons des politiques et/ou des procédures normales de choix de chemin, doivent être résolus de façon cohérente.

La règle qu'un SA n'annonce à ses SA voisins que les chemins qu'il utilise est fondamentale pour BGP. Cette règle reflète le paradigme de l'acheminement "bond par bond" généralement utilisé par l'Internet actuel.

6. Sélection du chemin avec BGP

Une des tâches majeures d'un locuteur BGP est d'évaluer les différents chemins de lui-même à un ensemble de destinations couvertes par un préfixe d'adresse, de choisir le meilleur, d'appliquer les contraintes de politique appropriées, puis de l'annoncer à tous ses voisins BGP. La question clé est la façon dont les différents chemins sont évalués et comparés. Dans les protocoles traditionnels de vecteur de distance (par exemple, RIP) il n'y a qu'une seule métrique (par exemple, le compte de bonds) associée à un chemin. À ce titre, la comparaison des différents chemins se réduit à la simple comparaison de deux nombres. Une complication dans l'acheminement inter SA provient de l'absence d'une métrique universellement acceptée parmi les SA qui puisse être utilisée pour évaluer les chemins externes. Chaque SA a plutôt son propre ensemble de critères pour les évaluations de chemin.

Un locuteur BGP construit une base de données d'acheminement qui comporte l'ensemble de tous les chemins praticables et la liste des destinations (exprimées par des préfixes d'adresses) accessibles à travers chaque chemin. Pour les besoins de la précision de l'exposé, il est utile de considérer l'ensemble des chemins praticables pour un ensemble de destinations associées à un préfixe d'adresse donné. Dans la plupart des cas, on s'attend à trouver seulement un chemin praticable. Cependant, lorsque ce n'est pas le cas, tous les chemins praticables devraient être conservés, et leur conservation accélère l'adaptation à la perte du chemin principal. Seul sera annoncé le chemin principal à chaque instant.

Le processus de choix de chemin peut être formalisé par la définition d'un ordre complet sur l'ensemble de tous les chemins praticables vers un ensemble de destinations associées à un préfixe d'adresse donné. Une façon de définir cet ordre complet est de définir une fonction qui transpose chaque chemin complet de SA en un entier non négatif qui note le degré de préférence du chemin. Le choix du chemin se réduit alors à appliquer cette fonction à tous les chemins praticables et à choisir celui qui a le plus fort degré de préférence.

Dans les mises en œuvre réelles de BGP, le critère d'allocation du degré de préférences à un chemin est spécifié comme information de configuration.

Le processus d'allocation d'un degré de préférence à un chemin peut se fonder sur plusieurs sources d'informations :

1. Informations explicitement présentes dans le chemin complet de SA.

2. Combinaison des informations qui peuvent être déduites du chemin complet de SA et d'informations en dehors de la portée de BGP (par exemple, des contraintes de politique d'acheminement fournies comme informations de configuration).

Les critères possibles pour l'allocation d'un degré de préférence à un chemin sont :

- Le compte de SA. Les chemins avec un plus petit compte de SA sont généralement meilleurs.

- Considérations de politique. BGP prend en charge des acheminements fondés sur une politique sur la base de la distribution contrôlée des informations d'acheminement. Un locuteur BGP peut connaître des contraintes de politique (aussi bien au sein de son propre SA qu'au dehors) et faire le choix de chemin approprié. Les chemins qui ne se conforment pas aux exigences de la politique ne sont pas examinés plus avant.

- La présence ou l'absence d'un ou de certains SA dans le chemin. Au moyen d'informations en dehors du domaine d'application de BGP, un SA peut connaître certaines caractéristiques de performances (par exemple, la bande passante, la MTU, le diamètre intra SA) de certains SA et peut essayer de les éviter ou de les préférer.

- L'origine du chemin. Un chemin appris entièrement par BGP (c'est-à-dire, dont les points d'extrémité sont internes au dernier SA sur le chemin) est généralement meilleur que celui pour lequel une partie du chemin a été apprise via EGP ou quelque autre moyen.

- Les sous-ensembles de chemin de SA. Un chemin de SA qui est un sous ensemble d'un plus long chemin de SA vers la même destination devrait être préféré à un plus long chemin. Tout problème dans le plus court chemin (comme une panne) sera aussi un problème dans le plus long chemin.

- La dynamique des liaisons. Les chemins stables devraient être préférés aux instables. Noter que ce critère doit être utilisé de façon très prudente pour éviter de causer des fluctuations de chemin inutiles. Généralement, tout critère qui dépend d'informations dynamiques peut causer une instabilité des chemins et devrait être traité avec une grande prudence.

7. Ensemble exigé de politiques d'acheminement prises en charge

Les politiques sont fournies à BGP sous la forme d'informations de configuration. Ces informations ne sont pas codées directement dans le protocole. Donc, BGP peut fournir la prise en charge de politiques d'acheminement très complexes. Cependant, il n'est pas exigé que toutes les mises en œuvre de BGP prennent en charge de telles politiques.

Nous n'essayerons pas de normaliser les politiques d'acheminement qui doivent être prises en charge dans chaque mise en œuvre de BGP ; nous encourageons vivement toutes les mises en œuvre à prendre en charge l'ensemble de politiques d'acheminement suivant :

1. Les mises en œuvre de BGP devraient permettre à un SA de contrôler les annonces de chemins appris par BGP aux SA adjacents. Les mises en œuvre devraient aussi prendre en charge un tel contrôle avec au moins la granularité d'un seul préfixe d'adresse. Les mises en œuvre devraient aussi prendre en charge un tel contrôle avec la granularité d'un système autonome, lorsque celui-ci peut être le système autonome qui a généré le chemin, ou le système autonome qui a annoncé le chemin au système local (système autonome adjacent). Il faut faire attention quand un locuteur BGP choisit un nouveau chemin qui ne peut pas être annoncé à un homologue externe particulier, alors que le chemin choisi précédemment était annoncé à cet homologue. Précisément, le système local doit expliquer explicitement à l'homologue que le chemin précédent est maintenant impraticable.

2. Les mises en œuvre de BGP devraient permettre à un SA de préférer un chemin particulier pour une destination (lorsque plus d'un chemin est disponible). Au minimum, une mise en œuvre devra prendre en charge cette fonctionnalité en permettant d'allouer administrativement un degré de préférence à un chemin sur la seule base de l'adresse IP du voisin d'où le chemin est reçu. La gamme permise des degrés de préférence alloués sera entre 0 et 2^(31) - 1.

3. Les mises en œuvre de BGP devraient permettre à un SA d'ignorer les chemins avec certains SA dans l'attribut de chemin AS_PATH. Une telle fonction peut être mise en œuvre en utilisant la technique décrite dans [RFC1104], et en

allouant "infini" comme "pondération" à de tels SA. Le processus de choix de chemin doit ignorer les chemins qui ont une "pondération" égale à "infini".

8. Interaction avec les autres protocoles d'acheminement extérieur

Les lignes directrices suggérées dans cette section sont cohérentes avec celles présentées dans [RFC1519].

Un SA devrait annoncer un agrégat minimal pour ses destinations internes par rapport à la quantité d'espace d'adresses qu'il utilise réellement. Cela peut être utilisé par les administrateurs de SA non BGP- 4 pour déterminer combien de chemins ouvrir à partir d'un seul agrégat.

Un chemin qui porte l'attribut de chemin ATOMIC_AGGREGATE ne doit pas être exporté dans BGP-3 ou EGP2, sauf si une telle exportation peut être accomplie sans briser les NLRI du chemin.

8.1 Échange d'informations avec EGP2

Le présent document suggère les lignes directrices suivantes pour les échanges d'informations d'acheminement entre BGP-4 et EGP2.

Pour ménager une migration en douceur, un locuteur BGP peut participer à EGP2, aussi bien qu'à BGP-4. Donc, un locuteur BGP peut recevoir des informations d'accessibilité IP au moyen d'EGP2 aussi bien qu'au moyen de BGP-4. Les informations reçues par EGP2 peuvent être injectées dans BGP-4 avec l'attribut de chemin ORIGIN réglé à 1. De même, les informations reçues via BGP-4 peuvent aussi bien être injectées dans EGP2. Dans ce dernier cas cependant, on a besoin de savoir qu'il peut se produire une éventuelle explosion d'informations lorsque un certain préfixe IP reçu de BGP-4 note un ensemble de réseaux consécutifs de classe A/B/C. L'injection de NLRI reçues de BGP-4 qui notent des sous-réseaux IP exige que le locuteur BGP injecte le réseau correspondant dans EGP2. Le système local va fournir les mécanismes pour contrôler l'échange des informations d'accessibilité entre EGP2 et BGP-4. Précisément, il est exigé d'une mise en œuvre conforme qu'elle prenne en charge toutes les options suivantes lors de l'injection des informations d'accessibilité reçues de BGP-4 dans EGP2 :

- injecter seulement par défaut (0.0.0.0); aucune exportation d'autres NLRI,
- permettre la désagrégation contrôlée, mais seulement de chemins spécifiques ; permettre l'exportation de NLRI non agrégées,
- ne permettre l'exportation que de NLRI non agrégées.

L'échange d'informations d'acheminement via EGP2 entre un locuteur BGP qui participe à BGP-4 et un pur locuteur EGP2 ne peut survenir qu'aux frontières du domaine (système autonome).

8.2 Échange d'informations avec BGP-30 n

Le présent document suggère les lignes directrices suivantes pour l'échange d'informations d'acheminement entre BGP-4 et BGP-3.

Pour ménager une migration en douceur, un locuteur BGP peut participer à BGP-3, aussi bien qu'à BGP-4. Donc, un locuteur BGP peut recevoir des informations d'accessibilité IP au moyen de BGP-3, aussi bien que de BGP-4.

Un locuteur BGP peut injecter les informations reçues par BGP-4 dans BGP-3 comme suit .

Si un attribut AS_PATH d'un chemin BGP-4 porte des segments de chemin AS_SET, l'attribut AS_PATH du chemin BGP-3 peut être construit en traitant les segments AS_SET comme des segments AS_SEQUENCE, d'où il résultera que le AS_PATH sera un seul AS_SEQUENCE. Bien que cette procédure perde des informations set/sequence, elle n'affecte pas la protection de suppression des acheminements en boucle, mais peut affecter les politiques, si celles-ci se fondent sur le contenu ou l'ordre de l'attribut AS_PATH.

Lors de l'injection de NLRI déduites de BGP-4 dans BGP-3, on doit être conscient de l'explosion potentielle d'informations lorsqu'un préfixe IP donné note un ensemble de réseaux consécutifs de classe A/B/C. L'injection de NLRI déduites de BGP-4 qui notent des sous-réseaux IP exige que le locuteur BGP injecte le réseau correspondant dans BGP-3. Le système local devra fournir des mécanismes pour contrôler l'échange d'informations d'acheminement entre BGP-3 et BGP-4.

Précisément, il est exigé d'une mise en œuvre conforme qu'elle prenne en charge les options suivantes lors de l'injection d'informations d'acheminement reçues de BGP-4 dans BGP-3 :
- injecter seulement par défaut (0.0.0.0), aucune exportation d'autres NLRI,
- permettre la désagrégation contrôlée, mais seulement de chemins spécifiques ; permettre l'exportation de NLRI non agrégées,
- ne permettre l'exportation que de NLRI non agrégées.

L'échange d'informations d'acheminement via BGP-3 entre un locuteur BGP qui participe à BGP-4 et un pur locuteur BGP-3 ne peut survenir qu'à la frontière du système autonome. Au sein d'un système autonome BGP, les conversations entre tous les locuteurs BGP de ce système autonome doivent être soit BGP-3, soit BGP-4, mais pas un mélange des deux.

9. Fonctionnement sur les circuits virtuels commutés

Lorsque on utilise BGP sur des sous-réseaux de circuits virtuels commutés (SVC, *Switched Virtual Circuit*) il peut être souhaitable de minimiser le trafic généré par BGP. Précisément, il peut être souhaitable d'éliminer le trafic associé aux messages périodiques KEEPALIVE *(garder en vie)*. BGP comporte un mécanisme pour fonctionner sur les services de circuits virtuels commutés (SVC) qui évite de garder les SVC ouverts en permanence et lui permet d'éliminer l'envoi périodique des messages KEEPALIVE.

La présente section décrit comment fonctionner sans messages KEEPALIVE périodiques pour minimiser l'usage des SVC lors de l'utilisation avec un gestionnaire de circuits SVC intelligent. Le schéma proposé peut aussi être utilisé sur des circuits "permanents", qui acceptent un dispositif du genre surveillance de qualité de liaison ou demande d'écho pour déterminer l'état de la connexité de la liaison.

Le mécanisme décrit dans cette section ne convient qu'entre locuteurs BGP qui sont directement connectés sur un circuit virtuel commun.

9.1 Établissement d'une connexion BGP

Le dispositif est sélectionné en spécifiant un Temps de garde de zéro dans le message OPEN.

9.2 Propriétés du gestionnaire de circuit

Le gestionnaire de circuit doit disposer de fonctionnalités suffisantes pour être capable de compenser le manque de messages KEEPALIVE périodiques :
- Il doit être capable de déterminer l'inaccessibilité de couche liaison dans une période finie prévisible de la survenance d'une défaillance.
- En déterminant l'inaccessibilité, il devrait :
  - lancer un temporisateur d'inactivité configurable (comparable à une valeur normale de temporisateur de garde),
  - essayer de rétablir la connexion de couche liaison.
- Si le temporisateur d'inactivité arrive à expiration, il devrait :
  - envoyer une indication de circuit interne DEAD à TCP.
- Si la connexion est rétablie, il devrait :
  - arrêter le temporisateur d'inactivité,
  - envoyer une indication de circuit interne UP à TCP.

9.3 Propriétés de TCP

Une petite modification doit être faite à TCP pour que le gestionnaire de circuit traite les notifications internes :
- DEAD : purger la file d'attente de transmission et interrompre la connexion TCP.
- UP : transmettre toutes les données en file d'attente ou permettre la poursuite d'un appel TCP sortant.

9.4 Propriétés combinées

Certaines mises en œuvre peuvent n'être pas capables de garantir que le processus BGP et le gestionnaire de circuit vont fonctionner comme une seule entité; c'est-à-dire qu'ils peuvent avoir des existences distinctes lorsque l'autre a été arrêté ou

a connu une défaillance.

Si c'est le cas, une interrogation périodique bidirectionnelle entre le processus BGP et le gestionnaire de circuit devrait être mise en œuvre. Si le processus BGP découvre que le gestionnaire de circuit est parti, il devrait fermer toutes les connexions TCP concernées. Si le gestionnaire de circuit découvre que le processus BGP est parti, il devrait fermer toutes ses connexions associées au processus BGP et rejeter toute autre connexion entrante ultérieure.

10. Conclusion

Le protocole BGP apporte un haut niveau de contrôle et de souplesse pour l'acheminement inter domaine tout en mettant en application les contraintes de politique et de performances et éviter les acheminements en boucle. Les lignes directrices présentées ici donnent un point de départ à l'utilisation de BGP pour fournir à la croissance de l'Internet un acheminement plus sophistiqué et gérable.

Appendice A Interaction de BGP et d'un IGP

Cette section présente les méthodes par lesquelles BGP peut échanger des informations d'acheminement avec un IGP. Les méthodes présentées ici ne sont pas proposées au titre de l'utilisation standard de BGP pour le moment. Ces méthodes sont présentées uniquement à des fins d'information. Les mises en œuvre peuvent vouloir retenir ces méthodes pour importer des informations d'IGP.

Ce sont des informations générales qui s'appliquent à tout IGP générique.

L'interaction entre BGP et un IGP spécifique sort du domaine d'application de cette section. Les méthodes pour les IGP spécifiques devraient être proposées dans des documents distincts. Les méthodes pour des IGP spécifiques pourraient être proposées pour un usage normalisé à l'avenir.

A.1 Vue générale

Par définition, tous les SA de transit doivent être capables de porter du trafic généré et/ou destiné à des sites localisés en dehors de ce SA. Cela exige un certain degré d'interaction et de coordination entre BGP et le protocole de passerelle intérieure (IGP, *Interior Gateway Protocol*) utilisé par ce SA particulier. En général, le trafic généré en dehors d'un SA donné va passer à travers des passerelles intérieures (des passerelles qui prennent en charge seulement l'IGP) et des passerelles frontières (qui prennent en charge à la fois l'IGP et BGP). Toutes les passerelles intérieures reçoivent des informations sur les chemins externes provenant d'une ou de plusieurs des passerelles frontières du SA via l'IGP.

Selon le mécanisme utilisé pour propager les informations de BGP au sein d'un SA donné, des soins particuliers doivent être pris pour s'assurer de la cohérence entre BGP et l'IGP, car des changements d'état vont vraisemblablement se propager à des vitesses différentes à travers le SA. Il peut y avoir une fenêtre temporelle entre le moment où un routeur frontière (A) reçoit de nouvelles informations d'acheminement de BGP qui ont été générées à partir d'un autre routeur frontière (B) au sein du même SA, et le moment où l'IGP au sein de ce SA est capable d'acheminer le trafic de transit vers ce routeur frontière (B). Durant cette fenêtre temporelle peut survenir un acheminement incorrect ou un "trou noir".

Afin de minimiser ces problèmes d'acheminement, le routeur frontière (A) ne devrait pas annoncer à ces homologues externes un chemin pour certains ensembles de destinations extérieures associées à un préfixe d'adresse X donnée via le routeur frontière (B) tant que les routeurs intérieurs au sein du SA ne sont pas prêts à acheminer le trafic destiné à ces destinations via le routeur frontière de sortie correct (B). En d'autres termes, l'acheminement intérieur devrait converger sur le bon routeur de sortie avant d'annoncer les chemins via ce routeur de sortie aux homologues externes.

A.2 Méthodes pour réaliser des interactions stables

L'exposé qui suit décrit plusieurs techniques capables de réaliser des interactions stables entre BGP et l'IGP au sein d'un système autonome.

A.2.1 Propagation des informations de BGP via l'IGP

Bien que BGP puisse fournir son propre mécanisme pour porter les informations de BGP au sein d'un AS, on peut aussi utiliser un IGP pour transporter ces informations, pour autant que l'IGP prenne en charge l'arrosage complet des informations d'acheminement (en fournissant le mécanisme pour distribuer les informations de BGP) et la convergence en un seul passage (rendant le mécanisme effectivement atomique). Si un IGP est utilisé pour porter les informations de BGP, la période de désynchronisation décrite précédemment ne va alors pas se produire du tout, car les informations de BGP se propagent au sein du SA en synchronisation avec l'IGP, et l'IGP converge plus ou moins simultanément avec l'arrivée des nouvelles informations d'acheminement. Noter que l'IGP ne porte que les informations de BGP et ne devrait pas interpréter ou traiter ces informations.

A.2.2 Protocole de passerelle intérieure étiquetée

Certains IGP peuvent étiqueter un chemin extérieur vers un SA avec l'identité de leurs points de sortie tout en les propageant au sein du SA. Chaque routeur frontière devrait utiliser des étiquettes identiques pour annoncer les informations d'acheminement extérieur (reçues via BGP) à la fois dans l'IGP et en propageant ces informations aux autres homologues internes (homologues au sein du même SA). Les étiquettes générées par un routeur frontière doivent identifier de façon univoque le routeur frontière en cause – les routeurs frontière différents doivent utiliser des étiquettes différentes.

Tous les routeurs frontières au sein d'un seul SA doivent observer les deux règles suivantes :

1. Les informations reçues d'un homologue interne par un routeur frontière A déclarant inaccessible un ensemble de destinations associées à un préfixe d'adresse donné doivent être immédiatement propagées à tous les homologues externes de A.

2. Les informations reçues d'un homologue interne par un routeur frontière A sur un ensemble de destinations accessibles associées à un préfixe d'adresses donné X ne peuvent pas être propagées à un homologue externe de A tant que A n'a pas un chemin IGP pour l'ensemble de destinations couvert par X et que les informations d'acheminement de l'IGP et de BGP n'ont pas des étiquettes identiques.

Ces règles garantissent qu'aucune information d'acheminement n'est annoncée en externe tant que l'IGP n'est pas capable de les prendre en charge correctement. Elles évitent aussi certaines causes de "trous noirs".

Une méthode possible d'étiquetage des chemins BGP et d'IGP au sein d'un SA est d'utiliser l'adresse IP du routeur frontière de sortie qui annonce le chemin extérieur dans le SA. Dans ce cas, le champ "gateway" dans le message BGP UPDATE est utilisé comme étiquette.

Une autre méthode d'étiquetage des chemins BGP et d'IGP est que BGP et l'IGP se mettent d'accord sur un identifiant de routeur. Dans ce cas, l'identifiant de routeur est disponible à tous les locuteurs BGP (version 3 ou au dessus). Comme cet identifiant est déjà univoque, il peut être utilisé directement comme étiquette dans l'IGP.

A.2.3 Encapsulation

L'encapsulation fournit le mécanisme le plus simple (en termes d'interaction entre l'IGP et BGP) pour porter le trafic de transit à travers le SA. Dans cette approche, le trafic de transit est encapsulé dans un datagramme IP adressé au routeur de sortie. La seule exigence imposée à l'IGP par cette approche est qu'il devrait être capable de prendre en charge l'acheminement entre les routeurs frontières au sein du même SA.

L'adresse du routeur de sortie A pour une destination extérieure X est spécifiée dans le champ d'identifiant BGP du message BGP OPEN reçu du routeur A (via BGP) par tous les autres routeurs frontière au sein du même SA. Afin d'acheminer le trafic à la destination X, chaque routeur frontière au sein du SA l'encapsule dans les datagrammes adressés au routeur A. Le routeur A effectue alors la désencapsulation et transmet les paquets originaux au routeur approprié dans un autre SA.

Comme l'encapsulation ne s'appuie pas sur IGP pour porter des informations d'acheminement extérieur, aucune synchronisation n'est requise entre BGP et l'IGP.

Des moyens d'identifier les datagrammes qui contiennent de l'IP encapsulé, du genre d'un code de type de protocole IP, devront être définis si cette méthode doit être utilisée.

Noter que, si un paquet à encapsuler a une longueur très proche de la MTU, ce paquet sera fragmenté par le routeur qui effectue l'encapsulation.

A.2.4 BGP omniprésent

Si tous les routeurs dans un SA sont des locuteurs BGP, il n'est alors pas nécessaire d'avoir d'interaction entre BGP et un IGP. Dans de tels cas, tous les routeurs dans le SA ont déjà des informations complètes sur tous les chemins BGP. L'IGP n'est alors utilisé que pour l'acheminement au sein du SA, et aucun chemin BGP n'est importé dans l'IGP.

Pour que les routeurs fonctionnent de cette façon, ils doivent être capables d'effectuer une recherche récurrente dans leur tableau d'acheminement. La première recherche va utiliser un chemin BGP pour établir le routeur de sortie, alors que la seconde recherche va déterminer le chemin d'IGP pour le routeur de sortie.

Comme l'IGP ne porte pas d'informations externes dans ce scénario, tous les routeurs dans le SA auront convergé aussitôt que tous les locuteurs BGP auront les nouvelles informations sur ce chemin. Comme il n'y a aucune nécessité de retarder la convergence de l'IGP, une mise en œuvre peut annoncer ces chemins sans autre délai dû à l'IGP.

A.2.5 Autres cas

Il peut y avoir des SA avec des IGP qui ne peuvent ni porter des informations BGP ni étiqueter les chemins extérieurs (par exemple, RIP). De plus, l'encapsulation peut être infaisable ou indésirable. Dans de telles situations, les deux règles suivantes doivent être observées :

1. Les informations reçues d'un homologue interne par un routeur frontière A qui déclare inaccessible une destination doivent être immédiatement propagées à tous les homologues externes de A.

2. Les informations reçues d'un homologue interne par un routeur frontière A sur une destination accessible X ne peuvent être propagées à aucun des homologues externes de A à moins que A ait un chemin IGP pour X et qu'un délai suffisant se soit écoulé pour que les chemins d'IGP aient convergé.

Les règles ci-dessus présentent des conditions nécessaires (mais pas suffisantes) pour la propagation des informations d'acheminement BGP aux autres SA. À la différence des IGP étiquetés, ces règles ne peuvent pas garantir que des chemins intérieurs vers le routeur de sortie approprié sont en place avant de propager les chemins aux autres SA.

Si le délai de convergence d'un IGP est inférieur à une petite valeur X, la fenêtre temporelle durant laquelle l'IGP et BGP ne sont pas synchronisés est aussi inférieure à X, et toute cette problématique peut être ignorée au prix de périodes transitoires (inférieures à X) d'instabilité d'acheminement. Une valeur raisonnable pour X est à l'étude, mais X devrait probablement être inférieur à une seconde.

Si le délai de convergence d'un IGP ne peut pas être ignoré, une approche différente est nécessaire. Des mécanismes et des techniques qui pourraient être appropriées dans cette situation feront l'objets d'études ultérieures.

Références

[RFC1104] H. Braun, "Modèles d'acheminement selon la politique", juin 1989.

[RFC1519] V. Fuller, T. Li, J. Yu et K. Varadhan, "Acheminement inter domaine sans classe (CIDR) : stratégie d'allocation et d'agrégation d'adresses", septembre 1993. *(D.S., rendue obsolète par la* RFC4632)

[RFC1771] Y. Rekhter, T. Li , "Protocole de routeur frontière v. 4 (BGP-4)", mars 1995. *(Obsolète, voir* RFC4271) *(D.S.)*

Considérations pour la sécurité

Les questions de sécurité ne sont pas abordées dans le présent mémoire.

Remerciements

Le présent document a été à l'origine publié comme RFC 1164 en juin 1990, dont les auteurs étaient Jeffrey C. Honig (Cornell University), Dave Katz (MERIT), Matt Mathis (PSC), Yakov Rekhter (IBM), et Jessica Yu (MERIT).

Les personnes suivantes ont aussi apporté des contributions essentielles à la RFC 1164 -- Guy Almes (ANS, puis Rice University), Kirk Lougheed (cisco Systems), Hans-Werner Braun (SDSC, puis à MERIT), et Sue Hares (MERIT).

Nous tenons à remercier explicitement Bob Braden (ISI) pour sa relecture de la précédente version de ce document.

La version mise à jour du document est le résultat des efforts du groupe de travail BGP de l'IETF avec Phill Gross (MCI) et Yakov Rekhter (IBM) comme éditeurs.

John Moy (Proteon) a contribué à la Section 7 "Ensemble exigé de politiques d'acheminement prises en charge".

Scott Brim (Cornell University) a contribué aux bases de la Section 8 "Interaction avec les autres protocoles d'acheminement extérieur".

La plus grande partie du texte de la Section 9 est la contribution de Gerry Meyer (Spider).
Des parties de l'introduction ont été reprises presque intégralement de [RFC1519].

Nous tenons à remercier Dan Long (NEARNET) et Tony Li (cisco Systems) de leur relecture et de leurs commentaires sur la version actuelle de ce document.

Le travail de Yakov Rekhter a été soutenu en partie par la National Science Foundation sous le contrat numéro NCR-9219216.

Adresse des auteurs

Yakov Rekhter
T.J. Watson Research Center IBM Corporation
P.O. Box 704, Office H3-D40
Yorktown Heights, NY 10598
téléphone : +1 914 784 7361
mél : yakov@watson.ibm.com

Phill Gross
MCI Data Services Division
2100 Reston Parkway, Room 6001,
Reston, VA 22091
téléphone : +1 703 715 7432
mél : 0006423401@mcimail.com

Liste de diffusion du groupe de travail IDR de l'IETF : bgp@ans.net
Pour s'y faire ajouter : bgp-request@ans.net