

---

# Vers des réflecteurs de routes plus intelligents

**Steve Uhlig, Cristel Pelsser, Bruno Quoitin, Olivier Bonaventure**

*Département d'ingénierie Informatique  
Université catholique de Louvain, Belgique*

---

*RÉSUMÉ. BGP est le protocole de routage interdomaine actuellement utilisé dans l'Internet. Au sein d'un système autonome, les routes interdomaines sont souvent distribuées par le biais de réflecteurs de routes (RR) BGP. Dans cet article, nous montrons qu'en ajoutant de l'intelligence dans les réflecteurs de routes, il est possible de fournir des services utiles dans les gros réseaux IP. Nous présentons d'abord un premier exemple d'utilisation des réflecteurs de routes intelligents pour réagir aux pannes de liens de peering. Le second exemple d'utilisation des réflecteurs de routes intelligents montre leurs bénéfices pour les VPN BGP/MPLS entre systèmes autonomes.*

*ABSTRACT. The Border Gateway Protocol (BGP) is the standard interdomain routing protocol in the Internet. Inside an Autonomous System (AS), the interdomain routes are often distributed by using BGP Route Reflectors (RR). We show that by adding intelligence to the RR, it is possible to provide useful services in large IP networks. As examples, we first show how a versatile RR can help an AS to better react to the failure of peering links. Our second example shows the benefits of using versatile RRs to inter-AS BGP/MPLS VPNs.*

*MOTS-CLÉS: routage interdomaine, BGP, BGP/MPLS VPN, ingénierie de trafic*

*KEYWORDS: interdomain routing, BGP, BGP/MPLS VPNs, traffic engineering*

---

## 1. Introduction

Le protocole de routage interdomaine BGP [REK 04] est utilisé actuellement par plus de 16000 systèmes autonomes afin d'échanger leurs routes interdomaine. La stabilité ainsi que la performance de BGP sont des éléments essentiels pour la stabilité et la performance de l'Internet global [FEL 04]. BGP est aussi utilisé par des fournisseurs d'accès Internet pour distribuer d'autres types d'informations comme les routes VPN BGP/MPLS [ROS 03].

Lorsqu'il est utilisé pour le routage interdomaine, BGP se base sur deux types de sessions établies au-dessus de connexions TCP. Deux routeurs BGP de systèmes autonomes distincts reliés par un lien physique utilisent une session eBGP pour échanger leurs routes interdomaines. Les routes interdomaines reçues par les routeurs de frontière d'un système autonome doivent être propagées au sein du système autonome. Ceci est fait d'habitude par le biais de sessions iBGP. La spécification originelle de BGP supposait qu'un graphe complet de sessions iBGP serait utilisé au sein du système autonome pour distribuer les routes interdomaines. Une conséquence de ce graphe complet est qu'un routeur BGP ne doit pas redistribuer sur une session iBGP une route qu'il a apprise d'une autre session iBGP. Le problème de ce graphe complet de sessions iBGP est que  $\frac{N \times (N-1)}{2}$  sessions iBGP sont nécessaires dans un système autonome ayant  $N$  routeurs BGP, ce qui devient assez vite ingérable dans les gros réseaux actuels qui peuvent comprendre plusieurs centaines de routeurs.

Deux solutions ont été proposées pour résoudre ce problème du graphe complet des sessions iBGP : les confédérations BGP [TRA 96] et les réflecteurs de routes [BAT 03]. Un réflecteur de route est un routeur BGP particulier, qui peut redistribuer sur des sessions iBGP les routes qu'il a apprises d'autres sessions iBGP. Un réflecteur de routes a deux types de voisins iBGP : ses voisins "clients" et ses voisins "non-clients". Typiquement, les voisins non-clients sont d'autres réflecteurs de routes. Un réflecteur de routes reçoit des routes de tous ses voisins iBGP et utilise son processus de décision BGP afin de déterminer les meilleures routes pour joindre chaque destination. Si la meilleure route a été reçue sur une session iBGP avec un voisin client, le réflecteur de route ré-annoncera cette route à tous ses voisins iBGP. Par contre, si la route a été reçue d'un voisin non-client, alors la route ne sera annoncée qu'aux voisins clients.

Dans la plupart des déploiements actuels des réflecteurs de routes [HAL 97], le but est de minimiser la charge CPU sur le réflecteur de route. Les réflecteurs de routes sont souvent considérés comme un moyen de résoudre le problème de la distribution des routes iBGP. Dans cet article, nous montrons qu'en exploitant l'information de routage disponible dans les réflecteurs de routes, il est possible de fournir de nouveaux services aussi bien à l'intérieur d'un système autonome qu'entre systèmes autonomes.

La structure de cet article est la suivante. Nous discutons d'abord dans la section 2 plusieurs limitations des réflecteurs de routes BGP. Ensuite, nous illustrons l'intérêt de réflecteurs de routes intelligents par deux exemples importants en pratique. Le premier exemple que nous décrivons à la section 3 montre l'utilisation d'un réflecteur de routes

intelligent qui permet à un système autonome de transit de répartir son trafic entre les liens avec ses systèmes autonomes voisins ainsi que de réagir aux pannes de liens de peering. Deuxièmement, nous montrons dans la section 4 comment les réflecteurs de routes peuvent aider pour l'établissement de Label Switched Paths (LSPs) MPLS interdomaines.

Notons que [FEA 04] propose une plate-forme de contrôle du routage (RCP) afin de contrôler la sélection des routes interdomaines à l'intérieur d'un système autonome ainsi qu'entre systèmes autonomes. Par rapport à [FEA 04], notre solution a plusieurs avantages. Premièrement, nous évitons les problèmes d'extensibilité et de fiabilité d'un large système centralisé en permettant une hiérarchie de réflecteurs de routes BGP. Deuxièmement, la réutilisation des réflecteurs de routes BGP nous permet de faire évoluer graduellement le routage interdomaine actuel sans nécessiter de nouveau protocole. En conséquence, notre approche peut être déployée de manière incrémentale dans un système autonome (ou une partie de celui-ci). De plus, nous allons plus loin que [BON 04] en montrant via des simulations comment des réflecteurs de routes intelligents peuvent aider les fournisseurs d'accès Internet à résoudre des problèmes opérationnels importants.

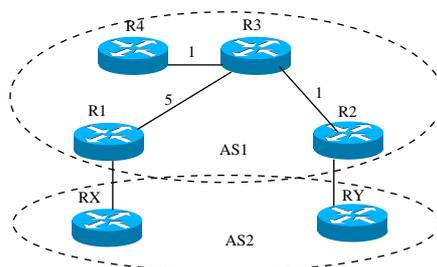
## 2. Limitations des réflecteurs de routes actuels

Les réflecteurs de routes actuels annoncent leur meilleure route à chacun de leurs voisins clients. Ceci permet au réflecteur de routes de calculer une seule meilleure route, mais ceci conduit à plusieurs problèmes. Le premier problème est que des boucles de routage et même des boucles de forwarding peuvent survenir lorsque les réflecteurs de routes sont utilisés. Plusieurs de ces problèmes ont été décrits dans la littérature [GRI 02a] et rapportés dans des réseaux réels [MCP 02]

Des extensions permettant à BGP d'annoncer plusieurs routes [WAL 02] ont été proposées pour résoudre ce problème, mais elles n'ont pas encore été implémentées ni déployées. A la place, les fournisseurs d'accès Internet utilisent des règles de bonne pratique lors de la conception des topologies iBGP [HAL 97, GRI 02b, XIA 03, BAT 03]. Ces règles de bonne pratique imposent des restrictions sur le graphe des sessions iBGP en fonction de la topologie intradomaine et de la localisation des réflecteurs de routes. En pratique, la topologie intradomaine change régulièrement lorsque les liens physiques et les routeurs tombent en panne ou lorsque des liens physiques sont ajoutés au réseau, ou même encore lorsque des outils d'ingénierie de trafic sont utilisés pour optimiser les poids intradomaine [FOR 02]. S'assurer que les conditions imposées par les règles de bonne pratique sont préservées après chaque changement intradomaine n'est pas trivial.

Si le CPU du réflecteur de routes n'est pas un goulot d'étranglement, une solution pour éviter les boucles de routage et de forwarding consiste à modifier le comportement des réflecteurs de routes. Au lieu de calculer sa propre meilleure route qui serait alors distribuée à tous ses clients, un réflecteur de routes pourrait calculer la meilleure

route qui serait calculée par chaque client si ce dernier avait la même information de routage BGP que le réflecteur de routes [MUS 04]. Etant donné qu'une des étapes du processus de décision BGP se base sur le poids intradomaine entre le routeur et le prochain saut de sa route BGP, le réflecteur de routes devrait connaître le poids intradomaine entre chacun de ses clients et chaque prochain saut BGP. Cette information peut être obtenue en calculant la table de routage intradomaine de chaque client ou en définissant un nouveau protocole afin de permettre à chaque client de fournir cette information à son réflecteur de route [MUS 04].



**Figure 1.** *Topologie de réseau simple*

Un autre problème qui peut survenir avec les réflecteurs de routes est le temps de convergence en cas de panne. Considérons la topologie de réseau du schéma 1. Supposons que AS2 est un fournisseur annonçant le préfixe  $P$  à partir des routeurs  $RX$  et  $RY$ . Considérons d'abord le cas d'un graphe complet de sessions iBGP dans AS1. Si le lien  $R2 - RY$  tombe en panne,  $R2$  supprimera sa route vers  $P$  de ses sessions iBGP et basculera immédiatement vers  $R1$  comme son prochain saut puisqu'il a reçu une route alternative de  $R1$  par iBGP.

Supposons à présent que  $R3$  soit un réflecteur de routes et qu'il préfère toutes les routes apprises par  $R2$  à cause du poids intradomaine. Avec le réflecteur de routes,  $R2$  n'apprendra pas les routes du routeur  $R1$  puisque  $R2$  annonce déjà la meilleure route. Lorsque le lien  $R2 - RY$  tombe en panne,  $R2$  ne connaît pas de route alternative et est obligé d'envoyer un message BGP WITHDRAW pour le préfixe  $P$  à  $R3$  pour forcer ce dernier à annoncer la route qu'il a apprise par  $R1$ . Pour une route unique, cet échange de messages BGP peut être rapide. Cependant, si ce n'est pas un préfixe mais 100000 routes que  $RX$  et  $RY$  annoncent, alors lorsque la session eBGP entre  $R2$  et  $RY$  tombe en panne,  $R2$  doit supprimer 100000 routes sur la session  $R2 - R3$ .  $R3$  devrait alors envoyer les 100000 routes apprises de  $R1$  sur la session iBGP entre  $R3$  et  $R2$ . Cela peut prendre un temps non-négligeable en fonction des performances du réflecteur de routes et de la structure des sessions iBGP dans le système autonome. Il est à noter que la taille actuelle des tables BGP dans Internet est de l'ordre de 140000 routes.

Avec les SLA strictes actuelles, il existe un besoin évident de réduire le temps de convergence du routage en cas de pannes. Un réflecteur de routes intelligent pour-

rait aider à le réduire en annonçant plusieurs routes vers un préfixe particulier à ses clients. En connaissant la table de routage intradomaine de chacun de ses clients, le réflecteur de routes intelligent peut déterminer aisément la meilleure route BGP, mais aussi la deuxième meilleure route que le client sélectionnerait si la première devenait injoignable. En utilisant les extensions à BGP proposées dans [WAL 02], le réflecteur de routes pourrait annoncer la meilleure et la deuxième meilleure route à chaque client. Cela assurerait que le client peut basculer très vite vers une nouvelle route lorsque la route primaire devient injoignable.

### 3. Réagir aux pannes de peering BGP

Comme mentionné dans [FEL 04], une large fraction des événements BGP concernent des peerings BGP entre systèmes autonomes, comme des pannes des liens de peering ou des événements de ré-initialisation de sessions eBGP. Ces événements peuvent avoir un impact majeur sur le flux du trafic et peuvent modifier la matrice de trafic. Idéalement, les routeurs BGP dans un système autonome devraient être capables de réagir à ces pannes de façon à ce que déplacement de trafic ne provoque pas de congestion dans d'autres parties du système autonome. A cette fin, se baser sur le routage BGP par défaut n'est pas suffisant car le choix de la meilleure route BGP ne dépend pas de la manière dont le trafic se distribue dans le système autonome. Les solutions traditionnelles d'ingénierie de trafic comme l'optimisation des poids intradomaine [FOR 02] ou l'utilisation de tunnels MPLS [AWD 02] ont des difficultés à s'adapter à de tels changements BGP car elles se basent uniquement sur la matrice de trafic intradomaine.

Dans cette section, nous présentons une manière d'utiliser un réflecteur de routes intelligent afin de réagir aux pannes des peering du réseau de GEANT (<http://www.geant.net>). Etant donnée la taille du réseau de GEANT, un seul réflecteur de routes est suffisant pour résoudre le problème, même si la solution que nous présentons pourrait être étendue à une hiérarchie de réflecteurs de routes. Pour des raisons de redondance, nous supposons que les routeurs de frontière du réseau de GEANT sont connectés en un graphe complet via des sessions iBGP, et que le réflecteur de routes a des sessions iBGP avec chaque routeur de frontière du réseau. De cette manière, même la panne du réflecteur de routes intelligent n'empêchera pas les routeurs BGP du réseau à avoir au moins une route pour joindre chaque destination. La fonction du réflecteur de routes intelligent est uniquement d'annoncer des messages iBGP aux routeurs de frontière afin d'influencer leur choix de la meilleure route pour joindre une destination particulière. Ceci peut être fait en manipulant les attributs des routes BGP, par exemple l'attribut `local-pref`.

Les principes de la solution sont les suivants. Premièrement, le réflecteur de routes collecte régulièrement des statistiques de trafic concernant les destinations les plus actives des routeurs de frontière du réseau. Ces statistiques peuvent être obtenues en utilisant des techniques décrites dans [LEI 04] ou [VAR 04]. Deuxièmement, le réflecteur de routes reçoit toutes les routes annoncées par les voisins du système autonome, par

exemple en utilisant l'extension à BGP proposée dans [WAL 02] qui permet aux routeurs BGP d'annoncer plusieurs routes pour chaque destination. Afin de contrôler le flux du trafic dans le système autonome, le réflecteur de routes contrôle les annonces iBGP qu'il envoie aux routeurs frontière du système autonome, ou à chaque réflecteur de routes présent dans chaque point de présence du système autonome, selon la configuration des sessions iBGP du réseau.

A partir de ces informations de routage et de trafic, le réflecteur de routes fait tourner régulièrement une heuristique évolutive [UHL 04a] afin d'améliorer la distribution du trafic dans le système autonome. Les détails de cette heuristique sont discutés en détail dans [UHL 04a]. Cette heuristique a déjà été appliquée au problème de l'ingénierie de trafic de systèmes autonomes non-transit [UHL 04b]. Nous avons adapté l'heuristique proposée dans [UHL 04a] pour le cas de systèmes autonomes de transit. Pour des raisons de limitation d'espace, nous ne décrivons pas les aspects techniques de cette heuristique dans ce papier. Cette heuristique étant un algorithme évolutif multi-objectifs, l'algorithme peut être utilisé pour tenir compte de différents objectifs comme répartir la charge du trafic entre les voisins BGP, réduire le coût total de la tarification du trafic, . . . Afin de mener à bien ces objectifs, l'algorithme sélectionne à chaque intervalle de temps pendant lequel l'algorithme tourne des changements iBGP afin d'influencer la meilleure route BGP utilisée par les routeurs d'entrée du réseau. Le coeur de l'algorithme est une recherche évolutive multi-objectifs [UHL 04a]. Cette heuristique multi-objectifs se base sur une population (ensemble de solutions candidates) d'individus, représentant chacun une solution potentielle au problème, i.e. un ensemble de messages iBGP à envoyer aux routeurs frontière. Chaque individu contient un ensemble de paires <routeur d'entrée, préfixe destination> qui définissent les modifications par rapport au routage par défaut BGP. Un individu est équivalent à un ensemble de modifications aux routes BGP faites par tous les routeur frontières. Pour trouver une solution au problème, l'heuristique utilise une recherche locale sur les individus et les mécanismes habituels de pression sélective propre aux algorithmes évolutifs [GOL 89]. A chaque intervalle de temps pendant lequel l'heuristique calcule les messages iBGP à envoyer, l'heuristique évolutive fait une itération sur la population afin d'améliorer la "qualité" des solutions potentielles en termes des multiples objectifs considérés comme décrit dans [UHL 04a]. La partie cruciale de l'heuristique afin d'obtenir de bonne solutions dans le cas multi-objectifs est que la sélection des solutions se fasse vis-à-vis de la Pareto-optimalité des solutions. Nous avons montré dans [UHL 04b] qu'un nombre très limité de messages iBGP par minute sont nécessaires dans le cas d'un système autonome non-transit. Le nombre de messages iBGP est du même ordre de grandeur dans le cas d'un système autonome de transit [UHL ], charge iBGP qui est bien inférieure à la charge normale de messages BGP dans un système autonome de transit [AGA 04].

GEANT est un réseau de communication pan-Européen multi-gigabit géré par DANTE et réservé pour une utilisation de recherche et d'éducation. GEANT a des interconnexions avec deux fournisseurs Internet, que nous appelons X et Y. GEANT a des peerings avec le système autonome X à quatre points différents et à deux différents points avec le système autonome Y, un des routeurs de GEANT ayant un peering avec

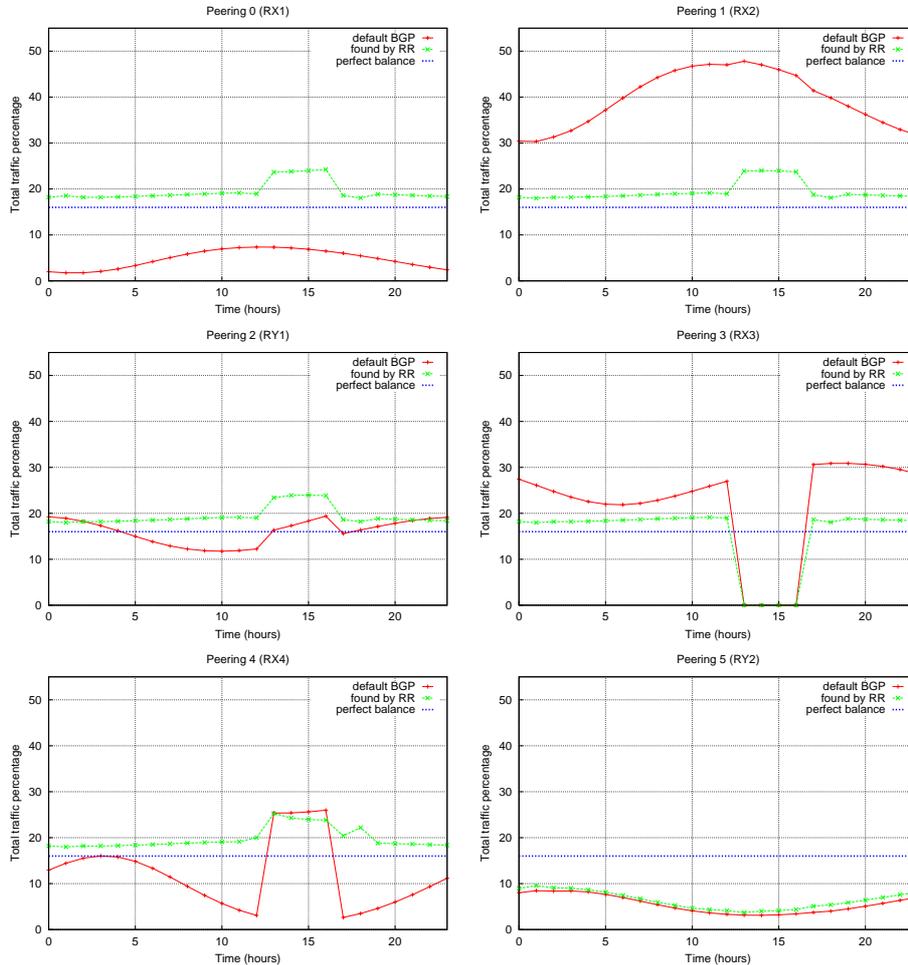
X et Y. Actuellement, GEANT utilise un graphe complet de sessions iBGP entre ses routeurs. En plus de la topologie de GEANT, nous avons obtenu une trace de tous les messages iBGP reçus par un routeur participant au graphe complet des sessions iBGP. Nous avons choisi dans cette trace une période d'un jour commençant à 9h41 le 8 Février 2004.

Nous avons aussi généré une demande de trafic artificielle comme pourrait l'être celle d'un réseau de transit. Nous avons sélectionné aléatoirement 1000 préfixes destinations comme étant les destinations les plus populaires du réseau de GEANT. Nous avons ensuite généré du trafic de chaque point d'entrée du réseau vers chacun de ces préfixes destinations selon une distribution de Weibull. La raison du choix de la distribution Weibull réside dans l'observation que l'essentiel du trafic dans l'Internet est échangé avec une fraction très limitée des préfixes joignables [FAN 99, UHL 02, REX 02]. Une distribution Weibull capture raisonnablement cette propriété. Après avoir généré le trafic total devant être envoyé de chaque point d'entrée dans le réseau vers chaque préfixe destination, nous avons simulé une évolution périodique du trafic pour chaque paire <point d'entrée, préfixe> durant la journée et avons ajouté une phase aléatoire (dépendant du point d'entrée) à cette évolution afin de simuler les différentes heures de pointe des sources de trafic situées à travers le monde. Plus de détails concernant cette trace synthétique de trafic peuvent être trouvés à [UHL 04c].

Le scénario de nos simulations consiste en du trafic synthétique décrit ci-dessus, que l'on suppose être reçu par chacun point d'entrée du réseau de GEANT et ayant pour destination les préfixes sélectionnés aléatoirement joignables via les voisins Internet du réseau.

Pour illustrer la manière dont un réflecteur de routes intelligent pourrait réagir à la panne d'un peering, nous avons développé un ensemble de scripts de simulation [UHL 04c] utilisant le simulateur CBGP [QUO 04]. Nos scripts simulent la panne d'un peering sur un routeur du réseau de GEANT, par des messages WITHDRAW envoyés par le routeur de frontière du réseau au réflecteur de routes intelligent (et aux autres routeurs iBGP). Le réflecteur de routes réagit en envoyant des messages iBGP aux routeurs de frontière afin de redistribuer la charge du trafic de manière aussi uniforme que possible sur les peerings Internet disponibles du réseau GEANT.

Chaque graphe de la figure 2 fournit l'évolution dans le temps du trafic pour chacun des six peerings externes (numérotés entre 0 et 5). Dans le titre de chaque graphe,  $RX?$  ( $RY?$ ) indique un peering avec le système autonome  $X$  ( $Y$ ). Chaque graphe contient trois courbes. Les courbes dont la légende est "default BGP" montrent l'évolution du trafic lorsque le choix de la meilleure route BGP par défaut est fait par BGP à chaque routeur d'entrée. Les courbes dont la légende est "found by RR" montrent l'évolution du trafic lorsque le réflecteur de routes intelligent influence le choix de la meilleure route afin de redistribuer la charge du trafic équitablement entre les peerings externes. Finalement, les courbes dont la légende est "perfect balance" représentent la situation irréaliste mais idéale où chaque peering verrait passer exactement un sixième du trafic total à chaque intervalle de temps, i.e. si la charge était répartie parfaitement entre les six peerings externes. Notons que la granularité temporelle à laquelle nous avons



**Figure 2.** Réaction du réflecteur de routes intelligent à la panne du peering 3 en  $t=13$ , le peering redevient fonctionnel en  $t=16$ .

simulé que le réflecteur de routes intelligent recalculait les messages iBGP à envoyer aux routeurs d'entrée est d'une heure. Nous avons montré dans [UHL 04b] qu'il était possible de travailler à des échelles de temps de quelques minutes, tout en maintenant une charge en termes de messages iBGP faible. Notons aussi que notre implémentation prototype du calcul des messages iBGP ne nécessite que quelques secondes sur un PC standard pour calculer les messages à annoncer aux routeurs frontière.

Sur la figure 2, nous montrons l'évolution du trafic pour chaque heure de la journée considérés sur les six peerings externes de GEANT. Nous avons intentionnellement espacé dans le temps les événements de la panne d'un peering et de son retour en

fonctionnement. Le peering numéro 3 est en panne à l'heure 13, et redevient fonctionnel à l'heure 16. Nous avons espacé la panne et le retour du peering afin de permettre de visualiser sur la figure 2 comment le trafic est modifié lorsque le peering tombe en panne.

La courbe "default BGP" sur la figure 2 indique que lorsque le choix par défaut du processus de décision BGP est utilisé, le trafic n'est pas idéalement réparti sur les peerings externes. Les peerings 1 et 3, en raison de leur position dans le réseau de GEANT, transportent une plus grande fraction de trafic que les autres peerings. Le peering 5 (avec le système autonome Y) transporte une faible quantité de trafic parce que ce peering se trouve sur le même routeur que le peering 1 (avec le système autonome X) et que les routes annoncées sur ce dernier peering sont typiquement meilleures que celles annoncées par le système autonome Y.

La figure 2 montre qu'en sélectionnant les routes iBGP, le réflecteur de routes intelligent est capable de forcer le trafic à se redistribuer de manière plus équitable entre les différents peerings externes. On peut le voir à travers la plus grande proximité des courbes "found by RR" de la distribution idéale du trafic ("perfect balance"). Utiliser un réflecteur de routes intelligent dans cette situation améliore la manière dont le trafic est réparti entre les peerings, ce qui permet de limiter la congestion sur les liens de peerings en cas de panne et améliore ainsi la performance globale vue par le trafic. Notons qu'ici nous avons considéré un objectif simple de répartir équitablement le trafic entre les peerings, en supposant que chaque lien de peering a une capacité égale. En pratique, les liens de peering peuvent avoir des capacités différentes ainsi que des coûts différents. Dans ce cas, il suffit de modifier la fonction objective que notre heuristique optimise. Un exemple du fonctionnement de notre heuristique dans un tel cas peut être trouvé dans [UHL 04b].

Lorsque le peering 3 tombe en panne à l'heure 13, l'ensemble du trafic qui aurait normalement été transporté sur ce peering est rerouté vers d'autres peerings externes. Avec le routage BGP par défaut, l'essentiel du trafic est alors redistribué vers le peering 4. Pour ce peering, l'augmentation en termes du pourcentage du trafic total (sur l'ensemble des peerings) est de 20 %. Lorsque le réflecteur de routes intelligent influence les routes des routeurs d'entrée, l'augmentation du trafic sur le peering 4 due à la panne du peering 3 est limitée à au plus 5 % du trafic total. Cela signifie que l'impact (en termes de redistribution du trafic) d'une panne d'un peering est significativement réduit. Une situation similaire survient lorsque le peering 3 redevient opérationnel à l'heure 16, avec la charge du trafic sur le peering 3 qui passe de 0 à près de 30 % du trafic total lorsque le routage BGP par défaut est utilisé. Avec le réflecteur de routes intelligent, l'impact sur la charge du trafic est limité et la répartition du trafic sur les peerings externes retourne à une situation similaire à celle avant la panne.

#### 4. Les réflecteurs de routes et MPLS

De nombreux gros fournisseurs d'accès Internet utilisent aujourd'hui MPLS pour fournir des services VPN BGP/MPLS à leurs clients commerciaux [ROS 03]. Aujourd'hui, ces services sont souvent fournis au sein d'un seul système autonome. Trois types de routeurs sont identifiés d'habitude pour ces VPN BGP/MPLS. Un routeur *CE* est un routeur qui appartient à un client et est géré par ce dernier. Un routeur *PE* est un routeur géré par le fournisseur et est directement connecté aux routeurs *CE*. Un routeur *PE* apprend typiquement les routes joignables par chaque routeur *CE* auquel il est connecté à travers une session spéciale IGP ou BGP [ROS 03]. Afin d'isoler les différents VPN, un routeur *PE* maintient une table de routage et de forwarding VPN (VRF) par VPN. BGP est utilisé par les routeurs *PE* pour distribuer le contenu de leurs VRF aux autres routeurs *PE* qui sont attachés aux mêmes clients VPN. Le forwarding des paquets VPN d'un *PE* à un autre se base sur l'utilisation de tunnels MPLS, GRE ou IPSec. Grâce à l'utilisation de ces tunnels, les routeurs du coeur, appelés routeurs *P*, n'ont pas besoin de maintenir de VRF par VPN. Comme BGP est utilisé pour distribuer les routes VPN dans le réseau, les réflecteurs de routes sont souvent utilisés pour éviter de recourir à un graphe complet de sessions iBGP entre les routeurs *PE*.

Un problème récurrent avec les VPN BGP/MPLS est que des sites VPN importants sont souvent attachés à deux routeurs *PE* différents. Ce double attachement est souvent nécessaire pour un objectif de redondance, mais une fois que deux liens existent, les clients exigent de pouvoir les utiliser aussi bien pour l'entrant que pour le sortant. Pour les paquets envoyés par le routeur *CE* vers le fournisseur, le *PE* utilisé dépend seulement de la configuration du réseau client. Pour les paquets envoyés par le fournisseur VPN vers le routeur *CE*, la possibilité de répartir la charge du trafic entre les deux routeurs *PE* dépend de la distribution des routes VPN aux routeurs *PE*. Une solution envisageable est l'utilisation d'identificateurs de routes par site VPN [ROS 03] pour s'assurer que chaque *PE* reçoit toutes les annonces de tous les routeurs *PE* attachés au même VPN. Cependant, cela augmenterait la taille des tables de routage BGP/MPLS qui sont déjà plus grosses que les tables BGP des routes Internet [NIC 04]. Un réflecteur de routes plus polyvalent pourrait être configuré pour annoncer une seule route lorsque l'extensibilité est importante, et plusieurs routes (en utilisant [WAL 02] par exemple) pour les sites VPN pour lesquels la répartition de charge est importante.

##### 4.1. VPN BGP/MPLS entre systèmes autonomes

L'étape suivante pour les VPN BGP/MPLS est de fournir ces services entre des systèmes autonomes distincts. Différentes solutions sont proposées dans [ROS 03]. Les fournisseurs d'accès Internet sont poussés par leurs clients à fournir des services VPN BGP/MPLS entre systèmes autonomes avec des SLA stricts [ZHA 03]. Pour fournir ces services, deux problèmes doivent être résolus. Premièrement, les routes

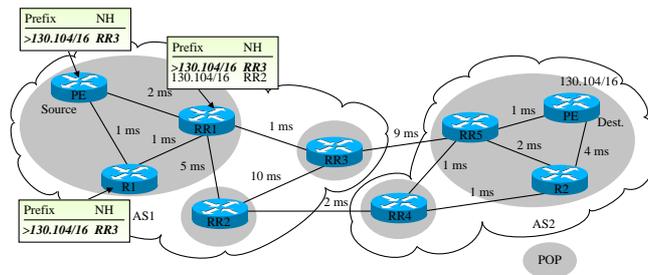
VPN doivent être distribuées entre les systèmes autonomes. Ceci peut être fait en connectant directement les réflecteurs de routes des systèmes autonomes qui coopèrent avec des sessions eBGP multi-hop afin de distribuer les routes VPN inter-fournisseurs [ROS 03]. Deuxièmement, des LSP interdomaines doivent être établis entre les routeurs *PE* pour transporter les paquets VPN encapsulés. Une solution pour établir ces LSP est d'utiliser RSVP-TE [AWD 01] pour établir des tunnels MPLS d'ingénierie de trafic entre routeurs *PE* distants. Comparé à LDP, l'avantage principal de RSVP-TE est qu'il permet de fournir des garanties de QoS et de protection aux tunnels. Différents moyens de protection ont été proposés dans la littérature [SHA 03]. En raison de limitations d'espace, nous considérons comme exemple dans cette section l'utilisation d'un LSP secondaire de bout en bout qui est noeud- et arc-disjoint du LSP primaire.

Dans un système autonome unique, lorsqu'un routeur de tête doit établir un tunnel d'ingénierie de trafic, celui-ci calcule son chemin explicite en se basant sur la topologie intradomaine. Le cas des LSP interdomaines est plus compliqué parce que le routeur de tête a une information limitée sur la topologie interdomaine. L'information "topologique" distribuée par BGP est limitée à la joignabilité. Pour chaque préfixe, un routeur BGP connaît le prochain saut vers ce préfixe et les systèmes autonomes à traverser pour le joindre. Il n'y a pas de métrique spécifique de qualité associée aux routes BGP ni d'information sur le chemin exact (niveau IP) pour joindre ce préfixe. En conséquence, le calcul de chemins interdomaines contraints en se basant sur l'information de routage interdomaine ne peut disposer que d'information topologique du système autonome local. Le calcul doit donc se faire de manière distribuée. Cependant, ces segments de chemin peuvent ne pas constituer un chemin de bout en bout qui satisfasse aux contraintes. Le calcul d'un tel chemin interdomaine requiert l'exploration de chemins alternatifs jusqu'à ce qu'un chemin qui respecte les contraintes soit trouvé.

De nombreux fournisseurs d'accès Internet utilisent les réflecteurs de routes pour des raisons d'extensibilité. Malheureusement, cette pratique a un coût en termes des routes disponibles dans les clients des réflecteurs de routes. La figure 3 illustre ce problème. Sur cette figure, nous considérons un routeur *PE* attaché au réseau 130.104.0.0/16 dans le système autonome AS2. Si un routeur *PE* dans le système autonome AS1 doit établir un LSP interdomaine pour joindre ce *PE*, il possède une seule route pour le joindre malgré que deux routes vers ce préfixe sont connues au sein du système autonome. Ceci est dû au fait que le réflecteur de route *RR1* sélectionne la meilleure route pour ses clients, et une seule route est annoncée aux clients. Nous pouvons donc observer que lorsque des réflecteurs de routes sont présents dans le système autonome, ces routeurs apprennent plus de routes que leurs clients. Les réflecteurs de routes sont donc prédisposés à calculer des chemins interdomaines contraints si l'information de joignabilité BGP doit être utilisée à cette fin.

Pour faciliter l'établissement de LSP interdomaines, nous proposons de fournir une fonctionnalité de calcul de chemins contraints dans les réflecteurs de routes. Nous proposons l'utilisation du protocole défini dans [VAS 04b] pour demander un calcul de

chemin à un réflecteur de routes et pour répondre à cette demande. Dans ce but, chaque routeur doit connaître les réflecteurs de routes qu'il peut contacter. Cette information peut être configurée manuellement ou dérivée des annonces intradomaines [VAS 04a]. Si un routeur a besoin d'un LSP interdomaine, il contacte un de ses réflecteurs de routes avec un message RSVP Path Request spécifiant les contraintes devant être respectées par le LSP. Le réflecteur de routes calcule un segment de chemin, dans le système autonome, à partir des routes apprises de ses voisins BGP et sa connaissance de la topologie intradomaine. Le segment de chemin calculé est alors renvoyé au routeur qui a initié la demande (la source PE dans la figure 3) dans un message Path Reply. L'initiateur de la demande place alors le segment de chemin calculé dans l'objet Explicit Route (ERO) et continue l'établissement du LSP le long du chemin spécifié par le ERO. A la réception d'un message Path, chaque routeur en aval sur le chemin du LSP qui doit compléter le ERO contacte un réflecteur de routes pour le calcul du segment de chemin.



**Figure 3.** Calcul du chemin par le réflecteur de routes

Si à un moment donné lors du calcul il s'avère impossible de compléter le chemin par rapport aux contraintes requises par le LSP, un message RSVP Path error est renvoyé vers la source du LSP. Le premier noeud intermédiaire qui a complété le ERO initie une autre requête Path request à un de ses réflecteurs de routes. Si le réflecteur de routes peut calculer un segment de chemin alternatif, le noeud tente un autre établissement le long de ce nouveau segment alternatif. Sinon, le message Path error est envoyé en amont. Ce mécanisme de cranckback [FAR 04] est nécessaire en raison de la quantité limitée d'information sur la qualité des routes interdomaines.

Afin d'évaluer cette utilisation de réflecteurs de routes intelligents, nous considérons un scénario où des VPN BGP/MPLS sont déployés dans deux systèmes autonomes de transit. Ce scénario est l'un des premiers besoins exprimés par les fournisseurs d'accès Internet pour des LSP interdomaines [ZHA 03]. Notre environnement de simulation contient deux systèmes autonomes de transit. Chaque système autonome contient plusieurs routeurs interconnectés dans des points de présence (POP). Un petit POP peut contenir un seul routeur alors qu'un gros POP peut être composé de plusieurs dizaines de routeurs. Les topologies de systèmes autonomes, avec les délais de liens et les routeurs groupés en POPs, sont celles collectées par le projet Rocketfuel

[MAH 02]. Nous avons assigné un débit de 10 Gbps pour chaque lien. Les systèmes autonomes sont interconnectés avec un lien de peering par ville où les deux systèmes autonomes ont chacun un POP. Chaque lien de peering entre systèmes autonomes a un délai de 1 msec. Pour établir des LSP interdomaines, nous considérons le cas de VPN inter-AS où chaque système autonome veut offrir des services de VPN vers les POP de l'autre système autonome. Pour cette raison, nous ajoutons un routeur *PE* dans chaque POP contenant plus d'un routeur. Ce routeur *PE* est connecté à deux routeurs différents dans le POP pour des raisons de redondance. Les sessions iBGP dans chaque système autonome sont configurées comme recommandé dans [BAT 03]. Un routeur dans chaque POP sert de réflecteur de routes, tous les routeurs dans le POP sont connectés via un graphe complet de sessions iBGP pour du routage intra-POP optimal. Les réflecteurs de routes sont aussi connectés entre-eux via un graphe complet de sessions iBGP. Les topologies détaillées peuvent être trouvées à [UHL 04c].

Nous considérons trois topologies. Leurs propriétés sont décrites dans les cinq premières colonnes du tableau 1. *Topo0* et *Topo1* ont plus ou moins le même nombre de liens intradomaines. Cependant, il y a une différence significative dans le nombre de liens de peerings. *Topo2* a un nombre élevé de liens par rapport aux autres topologies. La différence dans le nombre de liens et de noeuds est aussi plus grande pour cette topologie que pour *Topo0* et *Topo1*. Pour chaque topologie, nous avons utilisé C-BGP [QUO 04] pour calculer les tables de routage BGP de chaque routeur et réflecteurs de routes. Ensuite, nous avons essayé d'établir un graphe complet de LSP interdomaines d'ingénierie de trafic entre les routeurs *PE* dans chaque système autonome. Les contraintes considérées dans nos calculs sont le délai de bout en bout et le caractère noeud- et arc-disjoint sur les chemins. Chaque LSP est sujet à une contrainte de délai de bout en bout de maximum 100 ms. Pour chaque LSP primaire établi, nous établissons un LSP noeud- et arc-disjoint de bout en bout sujet aux mêmes contraintes que le LSP primaire. Ce LSP peut être utilisé pour la protection ou la répartition de charge.

Topologie	ASes		Noeuds	Liens		LSPs primaires		LSPs de backup	
	ASN1	ASN2		Intra	Inter	ASBR	RR	ASBR	RR
Topo0	AS3257	AS3967	521	554	3	100	100	0	64
Topo1	AS1755	AS3257	539	561	14	100	100	5	78
Topo2	AS1239	AS6461	948	1420	8	100	100	3	74

**Tableau 1.** *Topologies et pourcentage de LSPs interdomaines établis*

Le tableau 1 montre les résultats des simulations pour nos trois topologies interdomaines. Les quatre dernières colonnes du tableau 1 comparent le pourcentage de LSPs établis lorsque le calcul des chemins contraints se fait dans les routeurs frontières des systèmes autonomes (ASBR) et lorsque ce calcul est effectuée par les réflecteurs de routes (RR). Malgré la diversité dans les propriétés des topologies, nous obtenons des résultats similaires pour toutes les topologies. Tous les LSPs primaires ont pu être calculés indépendamment de l'endroit où ils ont été calculés. Cependant, nous observons

que les réflecteurs de routes fournissent un avantage lorsque l'on considère le calcul de chemins disjoints. Nos simulations montrent qu'un client d'un réflecteur de routes ne possède pas assez de routes pour pouvoir calculer deux chemins noeud- et arc-disjoint. Le nombre de LSPs secondaires établis est largement amélioré lorsque les réflecteurs de routes font le calcul.

## 5. Conclusions et extensions

Les réflecteurs de routes BGP ont été conçus pour résoudre le problème de l'extensibilité du graphe complet des sessions iBGP. Au lieu de servir simplement à la distribution des annonces iBGP, nous avons montré qu'en exploitant la connaissance de routage du réflecteur de routes, il est possible de fournir des services utiles. Nous avons fourni des résultats de simulations montrant deux de ces services. Le premier service est de permettre au réflecteur de routes intelligent de sélectionner les messages iBGP à annoncer en se basant sur des statistiques de trafic collectées dans le réseau pour permettre à un système autonome de transit d'améliorer la répartition de son trafic. Par rapport aux solutions classiques basées sur des LSPs MPLS ou des modifications aux poids intradomaines, un avantage important d'un réflecteur de routes intelligent est de pouvoir adapter le routage aux pannes de liens de peering. Le deuxième service est l'utilisation de réflecteurs de routes intelligents pour calculer les chemins de LSPs interdomaines afin de supporter des VPN BGP/MPLS entre systèmes autonomes.

Les extensions à ce travail concernent l'exploration d'autres situations où les réflecteurs de routes intelligents seraient utiles. Un réflecteur de routes intelligent pourrait aider à répartir la charge et la récupération rapide en annonçant plusieurs routes à leurs clients. Le sous-ensemble des routes devant être annoncées par le réflecteur de routes doit être évalué, par exemple par simulations. Un réflecteur de routes intelligent pourrait s'avérer l'endroit idéal pour surveiller les routes et détecter différents types d'attaques ou d'erreurs de configurations. Cela devrait être évalué par une implémentation prototype. Un autre exemple sont les extensions à BGP comme S-BGP ou SoBGP qui permettent aux routeurs de signer leurs messages BGP. Vérifier ces signatures dans chaque routeur BGP est une opération coûteuse qui pourrait être mieux effectuée par les réflecteurs de routes intelligents. Etant donnée la croissance prolongée de la taille des tables de routage VPN BGP/MPLS, un réflecteur de routes intelligent pourrait agréger les routes de chaque VPN avant de les réannoncer. Ceci serait utile pour des VPN BGP/MPLS interdomaines. Finalement, le calcul des chemins pour les LSP interdomaines avec des contraintes de débit et de délai profiterait d'extensions QoS à BGP. Nous collaborons actuellement avec un gros fournisseur d'accès afin de développer ces extensions.

## Remerciements

Nous remercions Nicolas Simar (DANTE) et Tim Griffin (Intel research) pour l'information sur GEANT, les traces iBGP et les tables de routage. Steve Uhlig est

financé par le FNRS (Fonds National de la Recherche Scientifique, Belgique). Ce travail a été partiellement financé par le Gouvernement Wallon (DGTRE) dans le cadre du projet TOTEM (<http://totem.info.ucl.ac.be>)

## 6. Bibliographie

- [AGA 04] AGARWAL S., CHUAH C., BHATTACHARYYA S., DIOT C., « The Impact of BGP Dynamics on Intra-Domain Traffic », *Proc. of ACM SIGMETRICS*, June 2004.
- [AWD 01] AWDUCHE D., BERGER L., GAN D., LI T., SRINIVASAN V., SWALLOW G., « RSVP-TE : Extensions to RSVP for LSP Tunnels », RFC 3209, December 2001.
- [AWD 02] AWDUCHE D., CHIU A., ELWALID A., WIDJAJA I., XIAO X., « Overview and Principles of Internet Traffic Engineering », Internet Engineering Task Force, RFC3272, May 2002.
- [BAT 03] BATES T., CHANDRA R., CHEN E., « BGP Route Reflection - An Alternative to Full Mesh IBGP », Internet draft, draft-ietf-idr-rfc2796bis-01.txt, work in progress, November 2003.
- [BON 04] BONAVENTURE O., UHLIG S., QUOITIN B., « The case for more versatile BGP Route Reflectors », Internet draft, draft-bonaventure-bgp-route-reflectors-00.txt, work in progress, July 2004.
- [FAN 99] FANG W., PETERSON L., « Inter-AS traffic patterns and their implications », *IEEE Global Internet Symposium*, December 1999.
- [FAR 04] FARREL A., SATYANARAYANA A., IWATA A., FUJITA N., ASH G., MARSHALL S., « Crankback Signaling Extensions for MPLS Signaling », Internet Draft, draft-ietf-ccamp-crankback-02.txt, work in progress, July 2004.
- [FEA 04] FEAMSTER N., BALAKRISHNAN H., REXFORD J., SHAIKH A., VAN DER MERWE J., « The case for separating routing from routers », *FDNA'04 workshop*, August 2004.
- [FEL 04] FELDMANN A., MAENNEL O., MAO M., BERGER A., MAGGS B., « Locating Internet Routing Instabilities », *ACM SIGCOMM2004*, August 2004.
- [FOR 02] FORT B., REXFORD J., THORUP M., « Traffic engineering with traditional IP routing protocols », *IEEE Communications Magazine*, , 2002.
- [GOL 89] GOLDBERG D., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.
- [GRI 02a] GRIFFIN T., WILFONG G., « Analysis of the MED oscillation problem in BGP », *ICNP 2002*, 2002.
- [GRI 02b] GRIFFIN T., WILFONG G., « On the correctness of iBGP configuration », *SIGCOMM'02*, Pittsburgh, PA, USA, August 2002, p. 17-29.
- [HAL 97] HALABI B., *Internet Routing Architectures*, Cisco Press, 1997.
- [LEI 04] LEINEN S., « Evaluation of Candidate Protocols for IP Flow Information Export (IPFIX) », Internet draft, draft-leinen-ipfix-eval-contrib-02, work in progress, January 2004.
- [MAH 02] MAHAJAN R., SPRING N., WETHERALL D., ANDERSON T., « Inferring Link Weights using End-to-End Measurements », *ACM SIGCOMM Internet Measurement Workshop*, November 2002.

- [MCP 02] MCPHERSON D., GILL V., WALTON D., RETANA A., « BGP persistent route oscillation condition », Internet draft, draft-ietf-idr-route-oscillation-01.txt, work in progress, 2002.
- [MUS 04] MUSUNURI R., COBB J., « A Complete Solution to Stable iBGP », *ICC'04*, 2004.
- [NIC 04] NICOLAS M., « BGP/MPLS VPN monitoring for troubleshooting, scalability verification and network migration safety », Presentation at MPLS2004, February 2004.
- [QUO 04] QUOITIN B., « C-BGP, an efficient BGP simulator », <http://cbgp.info.ucl.ac.be/>, March 2004.
- [REK 04] REKHTER Y., LI T., « A Border Gateway Protocol 4 (BGP-4) », Internet draft, draft-ietf-idr-bgp4-26.txt, work in progress, October 2004.
- [REX 02] REXFORD J., WANG J., XIAO Z., ZHANG Y., « BGP Routing Stability of Popular Destinations », *ACM SIGCOMM IMW'02*, November 2002.
- [ROS 03] ROSEN E., REKHTER Y., « BGP/MPLS IP VPNs », Internet draft, draft-ietf-l3vpn-rfc2547bis-01.txt, work in progress, September 2003.
- [SHA 03] SHARMA V., HELLSTRAND F., « Framework for Multi-Protocol Label Switching (MPLS)-based Recovery », Internet Engineering Task Force, RFC3469, February 2003.
- [TRA 96] TRAINA P., « Autonomous System Confederations for BGP », Internet RFC 1965, June 1996.
- [UHL ] UHLIG S., QUOITIN B., « BGP-based interdomain traffic engineering for transit ASes », Under submission, [http://cbgp.info.ucl.ac.be/apps.html#section\\_transit\\_te](http://cbgp.info.ucl.ac.be/apps.html#section_transit_te).
- [UHL 02] UHLIG S., BONAVENTURE O., « Implications of Interdomain Traffic Characteristics on Traffic Engineering », *European Transactions on Telecommunications*, , 2002.
- [UHL 04a] UHLIG S., « A multiple-objectives evolutionary perspective to interdomain traffic engineering in the Internet », *NIANT workshop in PPSN'04 conference*, September 2004.
- [UHL 04b] UHLIG S., BONAVENTURE O., « Designing BGP-based outbound traffic engineering techniques for stub ASes », *Comput. Commun. Rev.*, vol. 34, n° 5, 2004.
- [UHL 04c] UHLIG S., PELSSER C., QUOITIN B., « C-BGP simulation scripts and inter-AS topologies », available from <http://www.info.ucl.ac.be/~suh/cfip05>, August 2004.
- [VAR 04] VARGHESE G., ESTAN C., « The measurement manifesto », *Comput. Commun. Rev.*, vol. 34, n° 1, 2004, p. 9–14.
- [VAS 04a] VASSEUR J., AGGARWAL R., SHEN N., « IS-IS extensions for advertising router information », Internet draft, draft-vasseur-isis-caps-02.txt, work in progress, July 2004.
- [VAS 04b] VASSEUR J., (EDITORS) C. I., ZHANG R., VINET X., MATSUSHIMA S., ATLAS A., « RSVP Path computation request and reply messages », Internet draft, draft-vasseur-mpls-computation-rsvp-05.txt, work in progress, July 2004.
- [WAL 02] WALTON D., COOK D., RETANA A., SCUDDER J., « Advertisement of Multiple Paths in BGP », Internet draft, draft-walton-bgp-add-paths-01.txt, work in progress, November 2002.
- [XIA 03] XIAO L., WANG J., NAHRSTEDT K., « Reliability-aware iBGP Route Reflection Topology Design », *ICNP 2003*, November 2003.
- [ZHA 03] ZHANG R., VASSEUR J., « MPLS Inter-AS traffic engineering requirements », Internet draft, draft-ietf-tewg-interas-mpls-te-req-02.txt, work in progress, November 2003.