Optimization des réseaux de confédérations basés BGP

Mohamed Nassar

Optimisation des réseaux de confédérations basés BGP

MÉMOIRE

soutenu le 22 juin 2005

pour l'obtention du

DEA de l'Université Henri Poincaré – Nancy I

(Spécialité Informatique)

par

Nassar Mohamed el Baker

Composition du jury

Membres du jury :	Dominique Méry
	Didier Galmiche
	Noëlle Carbonell
	Olivier Festor
Encadrant :	Radu State

Laboratoire Lorrain de Recherche en Informatique et ses Applications - Equipe MADYNES

remerciements

Je tiens à remercier tout particulièrement :

- Monsieur Radu State, mon encadrant universitaire, pour son aide et ses conseils avisés,

- Monsieur Olivier Festor, responsable scientifique de l'équipe MADYNES, pour m'avoir accueilli au sein de son équipe et pour son soutien dans la réalisation de ce projet,

- l'ensemble de l'équipe MADYNES, pour l'ambiance conviviale qui y règne,

- le secrétariat du LORIA pour ses renseignements.

Table des matières

Table des figures

Introduction

Le protocole BGP (*Border Gateway Protocol*) est actuellement le protocole de facto du routage dans l'Internet. Plus que 10,000 domaines (*Autonomous Systems*) utilisent BGP pour échanger les informations d'accessibilité des adresses IP. Chaque domaine est une entité administrative indépendante sous contôle d'une organisation. IBGP (*Internal BGP*) est la variante de BGP utilisée par les routeurs BGP du même domaine. Le mécanisme de prévention des boucles de routage exige traditionnellement une session IBGP indépendante entre chaque paire de routeurs BGP du même domaine. La topologie virtuelle (*overlay network*) résultante s'appelle la maille complète (*full mesh*). Le passage à l'échelle de la maille complète provoque un problème sérieux dans les ASs à grande échelle. Les deux solutions existantes pour résoudre ce problème sont les réflecteurs de routes et les confédérations. La fiabilité des opérations IBGP dans les topologies alternatives est un critère important pour la conception d'une solution. Approfondir une approche théorique pour désigner une topologie virtuelle fiable et résiliente était la perspective de notre travail dans ce stage.

En particulier, les recommandations générales pour l'optimisation des confédérations ([13],[5]) sont insuffisantes et ne répondent pas à des questions essentielles. Une approche théorique pour l'évaluation et la conception des confédérations n'est pas encore formulée. La modélisation du réseau, la formulation du problème, et la proposition des solutions nous ont conduit à rechercher sur deux plans importants : le monde technique du protocole et les applications possibles de la théorie des graphes.

Le manuscrit est divisé en deux chapitres principaux. Dans le premier chapitre, nous commencerons par une brève présentation de l'architecture hiérarchique de l'Internet, du processus du routage, et du protocole BGP. Nous présenterons la problématique : le passage à l'échelle de IBGP et la difficulté de l'analyse de sa fiabilité, puis l'état de l'art sur l'optimisation des réflecteurs de routes et des confédérations.

Dans une seconde partie nous présenterons nos contributions : un modèle basé sur la théorie des graphes pour représenter le réseau, une métrique pour évaluer la fiabilité d'une topologie de confédération, une solution informatique pour résoudre le problème de l'optimisation dans un temps linéaire et la validation expérimentale de son exactitude.

EN annexe, nous donnerons un article en anglais que nous avons rédigé sur la même approche. Nous avons soumis cet article à la conférence IPOM 2005 et nous attendons actuellement les évaluations. IPOM est une conférence internationale IEEE qui s'intéresse à la gestion et à l'exploitation des réseaux IP.

Chapitre 1

Problématique et état de l'art

1.1 L'Internet et le routage

Aujourd'hui l'Internet rend service à la plus grande et la plus diverse communauté des utilisateurs dans le monde. Son infrastructure s'est transformée d'un réseau fédérateur (*core network*) qu'était NSFNET à une architecture plus distribuée opérée par des fournisseurs commerciaux comme UUNET, Qwest, Sprint, et des milliers des autres.

L'épine dorsale contemporaine de l'Internet est une collection de fournisseurs de services ayant des points de connexion appelés POPs (*Points of Presence*) répartis sur plusieurs régions. La collection des POPs et l'infrastructure qui les inter-connecte forment un réseau fournisseur. Les clients sont connectés aux fournisseurs par les facilités d'accès ou d'hébergement existantes dans un POP. Ces clients peuvent eux même être des fournisseurs. Pour permettre aux clients d'un fournisseur d'atteindre les clients d'un autre fournisseur, le trafic est échangé par des points d'accès publiques appelés NAPs (*Network Access Points*), ou par des interconnexions directes. Le terme ISPs (*Internet Service Providers*) est utilisé généralement pour référer n'importe qui fournissent le service de connexion à l'Internet, que ce soit directement à l'utilisateur final ou aux autres fournisseurs de service. Le terme NSP(*Network Service Provider*) est généralement utilisé pour décrire les fournisseurs de l'épine dorsale. Une caractérisation plus détaillée de la structure de l'Internet à partir de plusieurs points est publiée dans [9].

Les routeurs sont des dispositifs qui véhiculent le trafic entre les ordinateurs hôtes. Ils construisent des tables de routage qui contiennent les informations collectées sur les destinations accessibles et les meilleurs chemins vers ces destinations. Le routage basique est formé des étapes suivantes :

1. Les routeurs exécutent des programmes désignés par le nom de *protocoles* pour transmettre et recevoir les informations sur les routes.
2. Les routeurs utilisent ces informations pour remplir les tables de routage associées avec chaque protocole particulier.
3. Les routeurs balaient les tables de routage des différents protocoles (si plus qu'un protocole est exécuté) et sélectionnent le(s) meilleur(s) chemin(s) pour chaque destination.

4. Les routeurs associent à chaque destination l'adresse IP du dispositif du saut suivant (*next hop*) attaché par la couche de liaison de données (*data link layer* et l'interface locale de sortie qui va être utilisée lors de l'acheminement des paquets vers la destination. Notons que le dispositif du saut suivant peut être un autre routeur, ou la destination elle même.
5. Les informations d'acheminement vers le saut suivant (l'adresse de la couche liaison des données plus l'interface de sortie) sont placées dans la table d'acheminement du routeur.
6. Quand un routeur reçoit un paquet, il examine l'en-tête du paquet pour déterminer l'adresse de la destination.
7. Le routeur consulte la table d'acheminement pour obtenir l'interface de sortie et l'adresse du saut suivant pour arriver à la destination.
8. Le routeur exécute toute fonction additionnelle requise puis il envoie le paquet au dispositif approprié.
9. Cette procédure continue jusqu'à atteindre le hôte destinataire. Ce comportement reflète le paradigme du routage saut à saut utilisé généralement dans les réseaux à commutation de paquets.

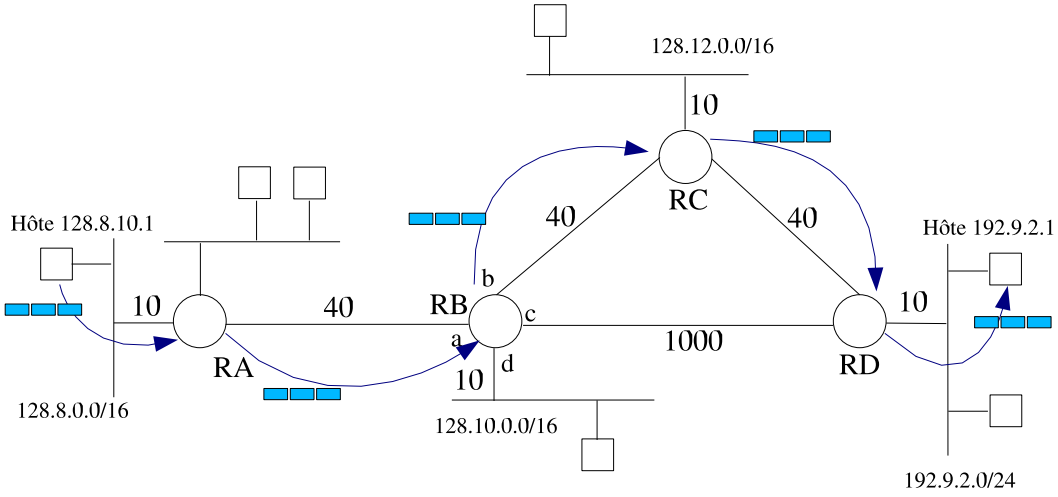

FIG. 1.1 – Comportement basique du routage

Dans la figure 1.1, un exemple simple est introduit pour montrer comment se déroule le routage. Les 4 routeurs RA, RB, RC et RD connectent quatre réseaux locaux (LANs) par des liaisons série. Les réseaux locaux ont les adresses IP suivantes : 128.8.0.0/16, 128.10.0.0/16, 128.12.0.0/16 et 192.9.2.0/24. Les routeurs ont normalement des adresses IP pour chaque interface mais ils sont désignés ici par leurs noms pour simplifier l'exemple. Chaque liaison série est assignée par une valeur qui indique le coût de transmission du trafic à travers cette liaison. Chacun des liens RA-RB, RB-RC et RC-RD a un coût de 40 tandis que le lien RB-RD a un coût de 1000. Pratiquement, le lien ayant un coût de 1000

Destination	Saut suivant	interface de sortie	Coût
128.8.0.0	RA	a	50
128.10.0.0	connecté	d	10
128.12.0.0	RC	b	50
192.9.2.0	RC	b	90

TAB. 1.1 – Table de routage de RB

peut être un fil téléphonique de débit 56 Kbps, et les liens ayant des coûts de 40 chacun peuvent être des liens de type T1 (débit=1.544 Mbps).

Les routeurs échangent leurs préfixes et construisent leurs tables de routage. Le tableau 1.1 donne une idée sur ce qu'est la table de routage du routeur RB. Pour donner un exemple sur le fonctionnement de routage de bout en bout, supposons que l'hôte 128.8.10.1 veut joindre le hôte 192.9.2.1. Il utilise d'abord la route par défaut installée manuellement pour envoyer le trafic à RA. RA balaye sa table de routage dans le but de trouver un réseau qui contient la destination et il va trouver que le réseau 192.9.2.0 est accessible à partir du RB comme saut suivant. Il y a deux routes de RB vers 192.9.2.0, mais dans un temps passé RB a décidé que la meilleure est celle avec comme saut suivant RC, de métrique 90, et d'interface b. Quand il reçoit le trafic, RD va trouver que l'hôte destinataire est directement connecté chez lui et il dirige le trafic à partir de l'interface correspondante.

1.2 Le protocole BGP

Les protocoles de routage sont divisés en deux grandes catégories : protocoles de routage intérieur IGP (*Interior Gateway Protocol*) et protocoles de routage extérieur EGP (*Exterior Gateway Protocol*). L'exemple précédent correspond à un protocole de routage IGP. Les IGPs, comme RIP(*Routing Information Protocol*) et OSPF(*Open Shortest Path First*) sont efficaces pour les réseaux de taille modérée, mais ils ne sont pas capables de supporter le routage avec plusieurs milliers des nœuds et centaines de milliers de routes. La structuration de l'Internet consiste à limiter le travail des IGPs dans des domaines administratifs indépendants ou ce qu'on appelle systèmes autonomes (*autonomous systems*) ASs. Un AS est en général sous contrôle d'une seule institution, comme par exemple une université ou une entreprise.

Un EGP est d'intérêt double : les routeurs dans un AS l'utilisent pour appliquer une politique de routage interne, et les routeurs dans des AS différents l'utilisent pour échanger les informations sur l'accessibilité des réseaux. Actuellement, le protocole BGP *Border Gateway Protocol* est le standard et unique protocole EGP pour le routage entre les domaines de l'Internet.

BGP utilise comme protocole de transport le TCP (*Transmission Control Protocol*). Les routeurs qui implantent BGP sont souvent nommés locuteurs (*BGP speakers*) ou entités BGP. Deux speakers qui établissent une connexion BGP pour échanger les tables de routages sont appelés voisins (*neighbors*) ou pairs (*peers*), et la connexion est appelée connexion de pair (*peer connection*) ou de voisinage (*neighboor conection*).

BGP est utilisé pour distribuer les informations d'accessibilité des préfixes parmi les ASs. Il est également utilisé extérieurement parmi les ASs et intérieurement à chaque AS. Une session BGP établie entre deux routeurs dans le même AS est désignée par *Internal BGP* ou IBGP, et une session entre deux routeurs qui résident dans deux ASs différents est désignée par *external BGP* ou EBGP. Les routeurs initiant des sessions EBGP sont appelés routeurs de bord (*border routers*). Ils jouent le rôle des portes entrée/sortie pour leurs ASs. Une session IBGP peut être établie par une connexion à sauts multiples grâce à TCP. En cas de EBGP standard, les deux pairs ont la restriction qu'ils doivent être connectés par un segment physique.

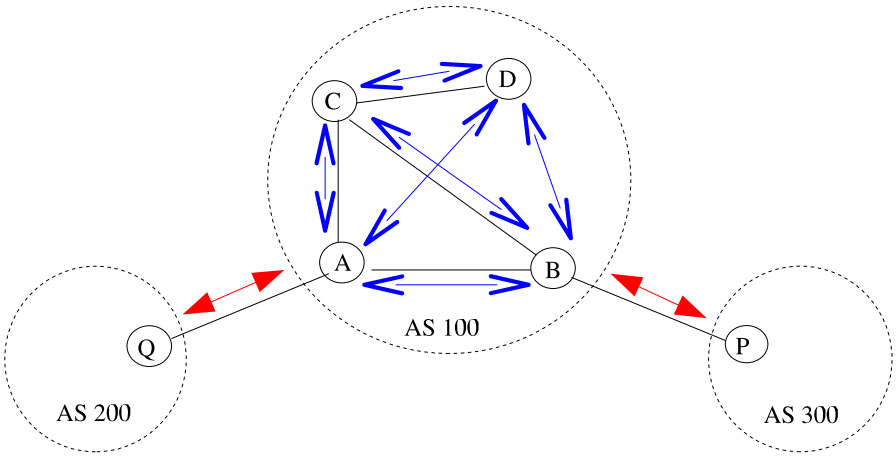

FIG. 1.2 – IBGP and EBGP

La figure 1.2 illustre un exemple de trois ASs. Les nœuds représentent les entités BGP et les lignes solides représentent les liens physiques. Nous avons deux sessions EBGP entre A et Q et entre B et P qui sont des routeurs de bords, et six sessions IBGP qui forment la maille complète de IBGP dans le AS 100. Les routeurs de bords A et B informent tous les locuteurs de leur domaine (par IBGP) sur les adresses des réseaux extérieurs accessibles (apprises par EBGP). Par l'intermédiaire de A et B, AS 100 peut annoncer ses propres préfixes, ou annoncer à AS200 un préfixe appris par AS300 et inversement. Remarquons que malgré que B et D n'aient pas de lien *physique*, ils arrivent à établir une session *logique*.

Les détails techniques du protocole sont hors du champs de ce manuscrit bien qu'ils forment une base importante pour comprendre les problèmes du passage à l'échelle et de fiabilité qui font objet de notre activité de recherche. Brièvement, les paquets utilisés dans une session BGP sont de quatre types : OPEN pour lancer la session, UPDATE pour échanger les mises à jour , NOTIFICATION pour fermer la session et KEEPALIVE pour maintenir la session en vie quand il n'y a pas des mises à jour en jeu. Le processus de sélection du chemin en un routeur BGP est basé sur les attributs BGP. Le routeur BGP déclenche ce processus quand il reçoit plusieurs chemins vers la même destination. Dans un message de mise à jour, chaque destination est associée avec des valeurs pour les différents attributs. Les plus importants sont AS-PATH, NEXT-HOP, MED, et LOCAL-PREF. Le routeur ignore la route si le NEXT-HOP est inaccessible (par IGP). Autrement,

il choisit celle ayant la plus grande préférence locale (LOCAL-PREF). En cas d'égalité, celle ayant le plus court chemin en nombre de domaines traversés (AS-PATH), puis celle ayant le plus petit discriminant de sorties multiples (MED) si les deux routes sont reçues à travers le même AS.

Les routeurs BGP détectent l'échec d'une session en utilisant un mécanisme de *time out*. Chaque routeur BGP maintient un temporisateur d'entretien (*KeepAlive Timer*) et un temporisateur d'attente *Hold Timer* pour chaque session BGP possédée par lui. A chaque expiration du temporisateur KeepAlive, il envoie un message KEEPALIVE vers le routeur pair associé avec cette session. En recevant un message KEEPALIVE ou tout autre message BGP, il met à zéro le temporisateur Hold. Une fois le temporisateur Hold expire, il assume que le routeur pair ne peut pas communiquer correctement et il ferme la session. Normalement, le routeur BGP attend de son pair au moins un message avant que le temporisateur Hold expire. Le délai ou la perte des messages aboutissent au dépassement du temporisateur Hold et parsuite à la ré-initialisation de la session. Pour plus d'informations sur le fonctionnement de BGP, le lecteur est invité à [5].

1.3 Problèmes : passage à l'échelle et fiabilité de IBGP

BGP est un protocole à vecteur de chemin. En d'autres termes, il associe à chaque destination la séquence des numéros des ASs identifiant le chemin qui mène à cette destination (l'attribut AS-PATH). Ceci est intéressant pour détecter les boucles de routage entre les ASs. En effet, chaque parleur EBGP porte le numéro d'AS où il appartient et néglige toute mise à jour pour une destination reçue par un autre parleur EBGP si le chemin associé avec cette destination contient son propre numéro d'AS. Ce mécanisme ne peut pas être appliqué à l'intérieur d'un AS où tous les routeurs ont le même numéro d'AS. La règle appliquée au routage intérieur énonce qu'une route apprise par l'intermédiaire de IBGP ne sera jamais annoncée une autre fois par l'intermédiaire de IBGP. Le seul moyen pour échanger totalement les informations du routage dans un AS est que chaque parleur maintient des sessions IBGP avec tous les locuteurs de son domaine. Cette conception est dite maille complète de IBGP.

Le nombre de sessions nécessaires pour former cette maille entre n routeurs est $n \times (n-1)/2$, et chaque routeur doit gérer $n-1$ sessions. Imaginons que pour 100 routeurs, le nombre requis de sessions est 4950. La conception de la maille complète à une grande échelle conduit à un grand nombre de sessions IBGP en même temps qu'à une grande consommation des ressources par routeur.

La gestion de routage dans les systèmes autonomes qui comptent des centaines de nœuds pose un problème sérieux pour les administrateurs des réseaux. Dans les ASs fournisseurs de services (ISPs) et les ASs de l'épine dorsale (*Backbone networks*), la majorité des routeurs implantent BGP. La conception de la maille complète de IBGP entre eux arrive rapidement à être hors contrôle. Du côté des ASs clients (*Customers*), même si un nombre limité de routeurs implantent BGP, la charge du IGP peut croître au delà du contrôle de l'administrateur.

Deux solutions sont employées pour résoudre le problème du passage à l'échelle de IBGP : remplacer la maille complète par une topologie de réflexion des routes [1], ou une

topologie de confédération [10].

La qualité du routage dans l'Internet est fortement dépendante de la fiabilité et de la stabilité des opérations IBGP à l'intérieur des domaines. Une session IBGP (ou BGP en général) peut souffrir de différentes faiblesses dans les couches inférieures à TCP. Un échec physique instantané cause une perturbation de IGP qui peut prendre un temps long (relativement aux temporisateurs BGP) pour converger. La perte d'une session peut provoquer des va-et-vient des routes et des adresses inaccessibles. Les routes correspondantes à une session déclarée perdue doivent être supprimées des tables de routages. Les messages de mise à jour envoyés pour effacer ces entrées peuvent déclencher une quantité énorme de traitement par les processus des décisions. La stabilisation du routage dans le domaine devient coûteuse en terme du temps et des ressources.

L'analyse de la fiabilité des réseaux IBGP est difficile, parce que les sessions IBGP peuvent être corrélées entre elles par les supports physiques communs, et parce que l'échec d'une session après une panne dans son chemin IGP est un évènement de nature probabiliste [6]. En effet TCP essaye de sauver la session grâce au mécanisme de retransmission des paquets. Les auteurs de [6] tendent à augmenter la probabilité de sauver la session. Leur idée est de modifier légèrement le mécanisme de retransmission de TCP, ce qui ne paraît pas pratiquement réalisable. Cependant, la conception de la topologie logique a une grande influence sur la résilience et la fiabilité du routage. L'aptitude de l'administrateur à déployer la topologie logique convenable est intéressante. La migration d'un réseau vers une topologie de confédération ou de réflecteurs de routes doit avoir comme effet direct un gain dans la fiabilité et l'efficacité de IBGP. Pour les domaines où la maille complète exige un nombre limité de sessions et le problème du passage à l'échelle n'est pas posé, la re-configuration du réseau n'est pas une bonne idée. La maille complète nous assure la plus grande indépendance logique (chaque paire de routeurs ont une session logique indépendante, la dépendance physique peut toujours exister). Par exemple, si une session expire sans être corrélée par la cause physique avec d'autres sessions, seuls les routeurs d'extrémité perdent contact, ce qui est un dégât minimal. Les topologies alternatives à la maille complète (réflexion des routes et confédérations) sont caractérisées par une dépendance logique de sorte que l'expiration d'une session peut causer une isolation dans le réseau.

Produire un cadre et des métriques pour évaluer la fiabilité des topologies alternatives et produire une méthodologie pour trouver la topologie logique optimale par rapport à l'infrastructure physique existante est la perspective de notre activité.

1.4 Les réflecteurs des routes

1.4.1 Définitions et techniques

Cette méthode consiste à spécifier dans l'AS un ensemble de routeurs IBGP pour être des réflecteurs des routes. La règle de IBGP pour empêcher les boucles est étendue pour cet ensemble de routeurs. C'est à dire ils peuvent refléter des routes d'un pair IBGP à un autre pair IBGP. Cette structure classifie les routeurs en trois groupes :

– Réflecteurs des routes ;

– Clients des réflecteurs ;
– locuteurs IBGP réguliers ou non-clients.

Du point de vue d'un réflecteur, il y a des sessions IBGP initiées avec des clients (ses propres clients) et d'autres avec des non-clients (les autres réflecteurs et les non-clients). Il se comporte comme un parleur IBGP régulier, mais en même temps il peut refléter les routes entre ses clients et ses non-clients, ou d'un client à un autre. Dans cette structure la maille complète est exigée seulement dans l'ensemble des réflecteurs et des non-clients. Deux réflecteurs peuvent avoir les mêmes clients. Des sessions de redondance peuvent exister également entre les clients d'un même réflecteur. Considérons la topologie représentée dans la figure 1.3, qui montre trois ASs inter-connectés par deux sessions EBGP. Dans AS300, RR est un réflecteur de routes avec trois clients C1, C2 et C3. Les routeurs NC1 et NC2 sont non-clients et ils sont totalement maillés avec RR. Chacun des clients forme une seule session IBGP avec RR. Une session de redondance a lieu entre C1 et C2. En cas de panne du RR, C1 et C2 ne perdent pas le contact. Le nombre total de sessions IBGP est de 7. Si la maille complète est configurée au lieu de la structure de réflexion, il faut que le nombre total de sessions IBGP soit 15.

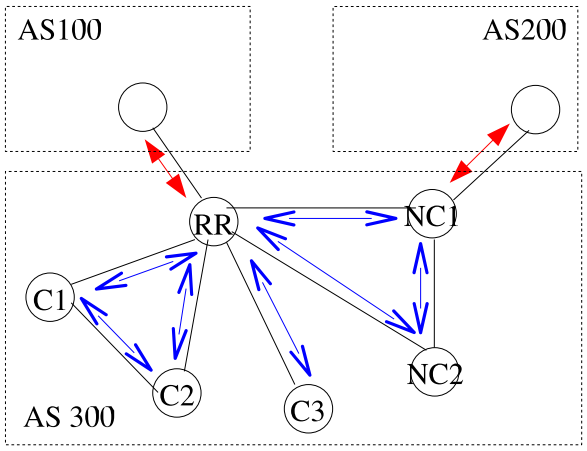

FIG. 1.3 – Composants d'une topologie de réflexion des routes

Nous pouvons définir plus clairement une réflexion comme une annonce de préfixes faite par un réflecteur d'un client à un autre client, ou d'un client à un non-client, ou d'un non client à un client. Les autres cas sont des annonces IBGP régulières et ne sont pas des réflexions. La figure 1.4 montre comment les règles de routage sont appliquées pour diffuser un préfixe parmi tous les locuteurs du domaine. Les trois cas sont représentés. Dans le cas (a) RR reçoit un préfixe de son pair EBGP, il l'annonce à travers toutes ses sessions IBGP pour clients et non-clients. Dans le cas (b), RR reçoit un préfixe de NC1, il le reflète à ses clients (mais pas à NC2 qui reçoit la route directement de NC1) et par EBGP au routeur de bord de l'AS100. Dans le cas (c) RR reçoit un préfixe de son client C1, il le reflète aux autres clients, aux non-clients et par EBGP.

La réflexion de routes fournit un autre avantage du point du vue passage à l'échelle : un réflecteur de routes réflète seulement le meilleur chemin pour chaque destination. Quand il reçoit plusieurs chemins vers la même destination, il applique premièrement le processus de

sélection du meilleur chemin, puis il reflète le résultat. Cette abstraction réduit le volume de mémoire nécessaire pour sauvegarder les informations de routage dans les routeurs de domaine en général, mais elle peut causer, lors d'une sélection incohérente entre le réflecteur et ses pairs, des boucles de routage et des pertes d'informations de routage. Pour maintenir une topologie conforme de BGP, les réflecteurs ne modifient pas certains attributs BGP durant la réflexion, comme NEXT-HOP, AS-PATH, LOCAL-PREF et MED.

Pour fournir un support de redondance pour la conception de réflexion, l'idée d'assembler les locuteurs dans plusieurs groupes ou *clusters* est introduite. Deux ou plusieurs réflecteurs peuvent servir les mêmes clients. Ces réflecteurs sont configurés avec le même identifiant de groupe ou *CLUSTER-ID*. La liste de groupes (*CLUSTER-LIST*) et l'identifiant d'origine (*ID-ORIGINATOR*) sont deux attributs additionnels pour empêcher les boucles des informations de routage dans un environnement de réflexion des routes.

La conception des réflecteurs de routes peut supporter une structure hiérarchique. Quand la maille complète entre les réflecteurs pose de nouveau la question du passage à l'échelle, une deuxième topologie de réflexion peut être basée sur la première. Pour les réseaux à grande échelle, une structure avec plusieurs niveaux de réflexion (pratiquement 2 ou 3) permet de réduire de plus en plus le nombre de sessions. Les réflecteurs d'un niveau bas sont en même temps des clients pour les réflecteurs du niveau juste en haut. La maille complète est exigée seulement pour le plus haut niveau.

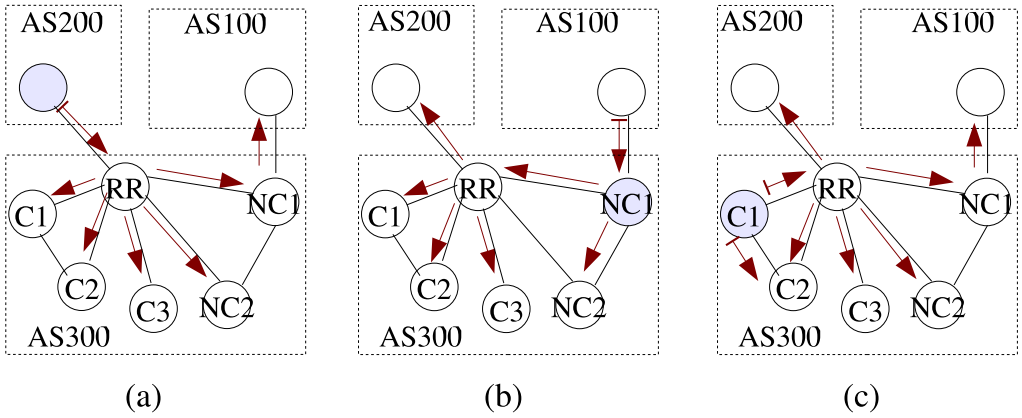

FIG. 1.4 – Règles pour les annonces dans une topologie de réflexion des routes

1.4.2 Optimisation de la topologie de réflexion

Influence de la topologie sur le travail de IBGP

La topologie logique de réflexion est formée par la division des routeurs en plusieurs clusters et la séléction dans chaque cluster du(ou des) routeur(s) qui va(vont) jouer le rôle de(s) réflecteur(s). Le choix de la topologie influence d'une manière directe sur la fiabilité du protocole intérieur. Par exemple la topologie en étoile avec un seul réflecteur n'est pas extensible et elle souffre d'un seul point de défaillance. En fait le réflecteur va

manipuler un grand nombre de sessions et va dépasser la capacité de ses ressources. Une fois qu'il tombe en panne, tout le processus BGP sera paralysé. Dans la figure 1.5, deux options différentes sont représentées pour déployer une topologie de réflexion au dessus de la structure physique de l'AS100. La conception (b) où B et F sont réflecteurs, A et D sont clients de C, et B et E sont clients de B ne constitue pas un bon choix. En effet, la session entre les réflecteurs B et C subit 4 sauts suivant le plus court chemin (F-B-E-D-C) et la longueur de toutes les sessions IBGP est 9 en terme du nombre des sauts, en la comparant avec 6 sauts pour toutes les sessions de (a). Désigner la réflexion suivant (a) est plus fiable parcequ'elle est plus superposable à la topologie physique. D'autre part, il faut choisir les composants les plus robustes pour supporter la réflexion. Si par exemple IGP indique que le lien A-E est moins fiable que les autres liens, un petit changement peut être effectué en (a) en choisissant D comme réflecteur au lieu de A. Le lien D-E sera utilisé au lieu de A-E ce qui assure un plus grand degré de fiabilité tout en conservant le même nombre de sauts pour l'ensemble des sessions (6).

Dans un réseau point de présence (POP), il y a des routeurs jouant le rôle du noyau qui sont choisis normalement comme des réflecteurs. Le problème de la conception est posé pour les réseaux étendus. La recommandation pratique est de tenter à superposer la topologie logique avec la topologie physique existante [13]. Xiao, Wang et Nahrstedt ont abordé l'approche théorique de ce problème dans [11], puis ils ont développé leur travail dans [12]. Ils proposent deux modèles des graphes non orientés pour représenter l' AS en question : un graphe physique $G(V, E)$ dont l'ensemble des sommets représente les routeurs et l'ensemble des arêtes représente les liens physiques, et un graphe logique (ou graphe de réflexion) $G_r(V_r, E_r)$. Dans le graphe logique, seuls les locuteurs IBGP forment l'ensemble des sommets, et les sessions IBGP forment l'ensemble des arêtes.

Chaque session est associée avec le chemin IGP correspondant dans le graphe physique. Le modèle logique est suffisamment flexible pour incorporer des configurations manuelles, de sorte que l'administrateur du domaine peut fixer quelques routeurs comme réflecteurs ou comme clients. Le modèle suppose l'existence d'un seul niveau de réflecteurs, et que chaque *cluster* ne doit pas comprendre plus d'un réflecteur, et qu'il n'y a pas de sessions de redondance entre les clients du même cluster. Ces suppositions simplifient le problème sans nuire à l'importance de l'approche.

Dans le paragraphe suivant, nous présentons les métriques proposées par ces deux articles pour l'évaluation de la fiabilité de la topologie de réflexion.

Métriques proposées pour l'évaluation d'une topologie

Dans [11], les auteurs définissent $\gamma(e)$ comme la fiabilité du lien physique e, qui signifie la probabilité d'un transfert avec succès d'un paquet à travers e, ou le pourcentage du temps où e travaille proprement. Un chemin IGP d'un nœud s à un nœud t est noté P_{st}. $P_{st} = (s, v1, v2....vn, t)$ où $(s, v1), (v1, v2)...., (vn, t)$ sont les liens physiques qui forment le chemin. La fiabilité d'un chemin P_{st} est $\gamma(P_{st}) = \prod_{e \in P_{st}} \gamma(e)$ ce qui est la probabilité d'un transfert avec succès d'un paquet le long du chemin.

Les auteurs dénotent le nombre des sauts (*hop count*) d'un chemin P_{st} comme $Hop(P_{st})$. Le nombre des sauts du graphe logique est la somme des nombres des sauts des chemins IGP de toutes les sessions : $Hop(G_r) = \sum_{s \in E_r} Hop(P_s)$, le nombre des sauts du graphe

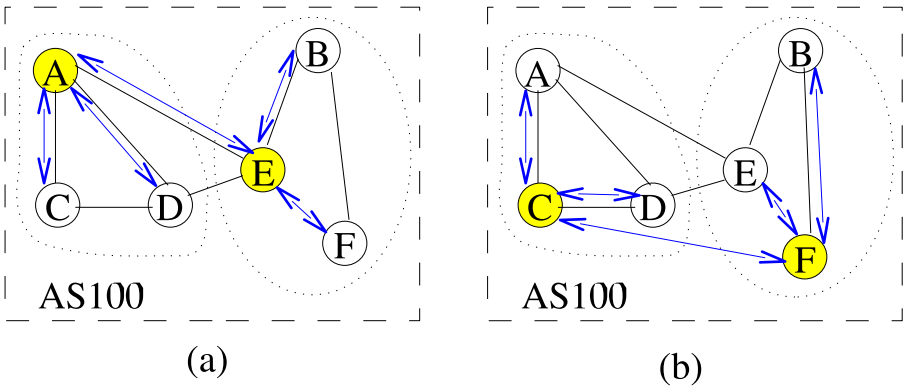

FIG. 1.5 – Deux options pour la conception d'une topologie

mesure *l'efficacité* de IBGP. Si la valeur est petite, le trafic du routage est faible et il prend une petite durée pour propager l'information à tous les locuteurs. D'autre part, ils définissent *la fiabilité* du graphe logique comme le produit des fiabilités de toutes les sessions : $\gamma(G_r) = \prod_{s \in E_r} \gamma(P_s)$.

Optimiser le graphe de réflexion consiste à trouver la topologie la plus efficace et en même temps la plus fiable pour IBGP. L'optimisation tend à minimiser le nombre de sauts $Hop(G_r)$ et à maximiser la fiabilité $\gamma(G_r)$. La seule contrainte mise pour la solution est que chaque parleur ne doit pas supporter plus que α sessions. Dans le but d'agréger les deux métriques, les auteurs utilisent la fonction logarithmique pour transformer la multiplication dans la formule de la fiabilité en une addition. Maximiser $\gamma(G_r)$ revient à minimiser la négation de son logarithme. La nouvelle métrique est à minimiser : $w(G_r) = \sum_{s \in E_r}[(-1)\theta \log(\gamma(P_s)) + (1-\theta)Hop(P_s))$, avec $\theta \in [0,1]$. Les auteurs choisissent θ proche de 1 parce qu'ils pensent que pratiquement la fiabilité est plus importante et utilisent l'efficacité pour casser l'égalité entre les topologies de même fiabilité.

Les mêmes auteurs proposent une nouvelle approche dans [12]. Le modèle physique est basé sur la fiabilité des composants élémentaires du réseau. La fiabilité d'un routeur est fonction de la puissance de son unité centrale de traitement (CPU), et sa capacité mémoire. Si le routeur manipule de façon concurrente un grand nombre de sessions, il va consommer une grande quantité de ses ressources, et il peut être surchargé ou hors mémoire. Ainsi, le modèle fixe une limite supérieure au nombre de sessions possédées c_i pour un routeur i.

Les auteurs assument que les évènements des échecs (incluant les encombrements intenses) pour les liens physiques sont conformes à un processus de Poisson, et dénotent la fréquence d'échec d'un lien physique l par w_l. De même, l'évènement d'échec d'un routeur i est un processus de Poisson et la fréquence d'échec est notée v_i. La session expire si l'un des deux routeurs d'extrémité tombe en panne, avec une probabilité p_r si un des autres routeurs de son chemin IGP tombe en panne, et avec une probabilité p_e si un lien physique de son chemin IGP tombe en panne. En réalité, dans les réseaux de l'épine dorsale, il est rare que deux composants physiques échouent simultanément. Ainsi, le modèle considère les différentes pannes comme des évènements indépendants.

Dans le modèle logique, les auteurs supposent que les routeurs sont tous des locuteurs

IBGP ($V = V_r$) (AS de transit) et qu'une simple extension peut ramener au cas général.

Les métriques proposées sont la durée de vie prévue (*Expected Life Time ELT*) et la perte des sessions prévue (*Expected Session Loss ESL*). Étant donné un réseau de réflexion, la durée de vie prévue ELT se rapporte au temps restant dès maintenant à la prochaine panne de IBGP. Basant sur les fréquences d'échec des éléments du réseau, la fréquence d'échec du réseau est calculée comme suit :

$$R_f(G_r) = \sum_{i \in V} v_i + \sum_{\substack{l \in E \\ k_l > 0}} w_l [1 - (1 - p_e)^{k_l}]$$

où k_l est le nombre de sessions supportées par l. La durée de vie prévue n'est autre que :

$$ELT(G_r) = \frac{1}{R_f(G_r)}$$

Soit m le nombre total des sessions IBGP. La perte des sessions prévue par unité de temps pour un routeur i est :

$$ESL(i) = \frac{h_i + p_r k_i}{m} v_i$$

- où h_i est le nombre de session possédées par i
- et k_i est le nombre de sessions dont leurs chemins IGP passent par i

De façon analogue, la perte de sessions prévue par unité de temps pour un lien physique l est :

$$ESL(l) = \frac{p_e k_l}{m} w_l$$

Enfin, la perte de sessions prévue pour le réseau entier $ESL(G_r)$ est par définition la perte maximale de tous les routeurs et les liens physiques :

$$ESL(G_r) = \max_{j \in V \cup E} ESL(j)$$

Optimiser le graphe de réflexion dans ce contexte revient à trouver la topologie logique caractérisée par un ELT maximum et un ESL minimum. Les deux métriques peuvent être agrégées ensemble comme : $\eta ESL(G_r) + \epsilon R_f(G_r)$ avec η est très grand par rapport à ϵ. Ce qui a pour effet lors de l'optimisation, de donner la priorité à réduire ESL, et parmi plusieurs topologies logiques avec ESL minimum, de casser l'égalité par R_f. Les auteurs prouvent que le problème d'optimisation est NP-difficile, et donnent des solutions heuristiques dont ils valident l'efficacité par des expériences informatiques. Un résultat important de ce travail est de montrer que le nombre optimal des réflecteurs ne doit être ni le minimum (topologie en étoile) ni le maximum (maille complète).

1.5 Les confédérations

1.5.1 Définitions et techniques

La confédération BGP est la deuxième solution pour remédier à l'explosion de la maille de IBGP dans un AS. Le concept de la confédération est qu'un grand système autonome

peut être divisé en un nombre des systèmes autonomes plus petits appelés AS secondaires (*sub-ASs*) ou membres AS (*AS members*). Chaque sub-AS a un numéro d' AS différent. La conséquence directe est que EBGP doit être le protocole de routage entre eux. Les sessions EBGP entre les sub-ASs sont appelées sessions EBGP intra-confédération (*intra-confederation EBGP sessions*), parce qu'elles sont un peu différentes des sessions EBGP régulières. La différence vient quand les préfixes sont échangés à travers les sessions : comme dans IBGP, les attributs BGP comme NEXT-HOP, MED, et LOCAL-PREF sont préservés, pourtant comme dans EBGP, l'attribut AS-PATH est modifié. AS-PATH joue son rôle habituel pour empêcher les boucles de routage entre les sub-ASs. A l'intérieur de chaque sub-AS, toutes les règles de IBGP sont appliquées. Par exemple, tous les locuteurs BGP à l'intérieur du sub-AS doivent être maillés entièrement. Une architecture de réflexion de routes peut être déployée s'il y en a besoin. La figure 1.6 montre une topologie d'une confédération. AS100 est divisé en deux sub-ASs : AS64520 et AS64530. Les sub-ASs sont protégés du monde extérieur et peuvent prendre n'importe quel numéro d'AS. Les numéros peuvent être choisis dans un espace privé (64512 à 65534) pour ne pas utiliser des numéros d'AS formels. Quand un sub-AS annonce un préfixe à un autre sub-AS, il ajoute son numéro d'AS dans un champ spécifique au début de l'AS-PATH. Quand un sub-AS annonce un préfixe à un pair EBGP extérieur, il enlève la séquence des sub-ASs de l'AS-PATH et il ajoute à sa place le numéro d'AS de la confédération. La confédération apparaît comme un simple AS et sa topologie reste invisible.

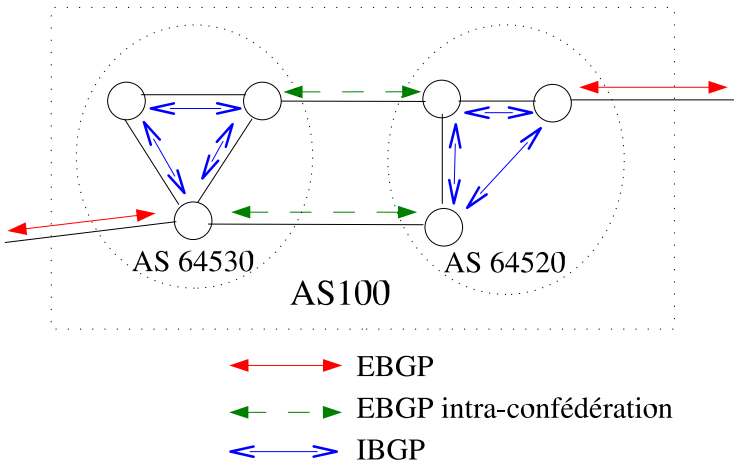

FIG. 1.6 – Éléments d'une confédération

L'avantage distingué des confédérations est que chaque sub-AS peut déployer un IGP indépendamment des autres sub-ASs. Leur inconvénient est que la migration d'un réseau pour être une confédération exige une re-configuration importante de tous les routeurs et un changement important dans la topologie logique. De plus, le routage à travers une confédération peut être sous optimal sans fixer manuellement les politiques de BGP. Par exemple prenons pour une même destination les deux chemins suivants : Chemin(1)=[AS100, AS200] et Chemin(2)=[AS300,AS400,AS200]. En BGP standard, le routeur doit choisir Chemin(1) parce'il est plus court. Cependant, AS100 est une confédéra-

tion, et en réalité Chemin(1)=[(AS64520, AS64530, AS64540), AS200]. Par conséquence, Chemin(2) est plus court et le routage est sous optimal. Le routage à l'intérieur de la confédération peut souffrir du même problème, parce que une séquence de sub-ASs n'influence pas sur la longueur d'un chemin. Par exemple du point de vue d'un sub-AS (AS64540), les deux chemins [(AS64520),AS300] et [(AS64530,AS64520),AS300] ont la même longueur, et AS64540 peut choisir l'un ou l'autre pour transmettre le trafic vers AS300. Le comportement du routage à l'intérieur d'une confédération peut être affecté en fixant des politiques additionnelles en fixant par exemple l'attribut LOCAL-PREF.

1.5.2 Architecture hub-and-spoke

Une bonne conception d'un réseau BGP doit satisfaire les propriétés suivantes : une complexité réduite, une simple procédure de routage, et en même temps une forte fiabilité. L'architecture en étoile (*hub-and-spoke architecture*) est recommandée dans la littérature ([5],[13]). Cette architecture a un sub-AS au centre qui joue le rôle du cœur du réseau. Les autres sub-ASs sont connectés seulement à lui et l'utilisent comme un central de transit. La figure 1.7 montre une architecture de hub-and-spoke.

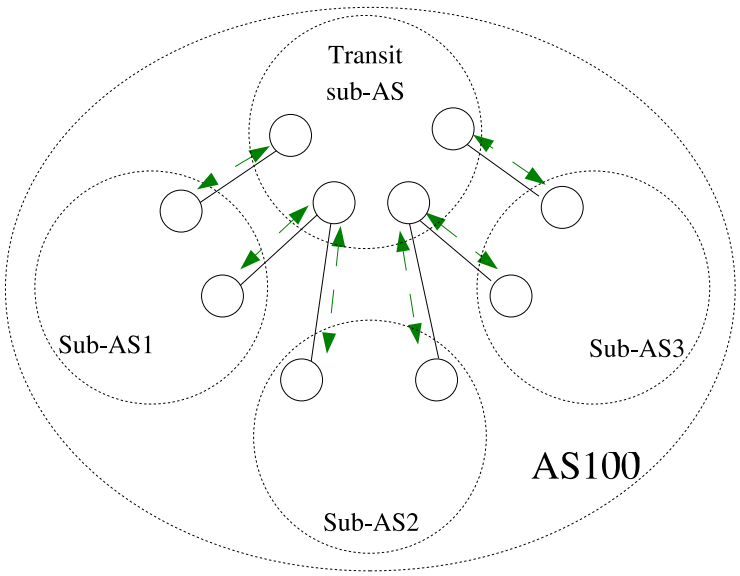

FIG. 1.7 – Architecture hub-and-spoke

Les avantages d'une telle architecture sont :
- un nombre réduit de sessions EBGP intra-confédérations ;
- un routage consistent et prédit entre les ASs. Le trafic d'un sub-AS non-transit à tout autre sub-AS non-transit subit toujours deux sauts AS.

Un nombre réduit de sessions EBGP intra-confédération est préféré, parce que si un sub-AS a plusieurs sessions EBGP intra-confédération, il peut recevoir plusieurs copies de la même route, ce qui résulte en un trafic redondant dans le réseau et un traitement redondant au niveau du routeur. Mais d'autre part la redondance augmente la résilience

du réseau en cas des défaillances des composants. Par exemple, si un sub-AS est connecté au sub-AS cœur par une seule session supportée par un lien physique. La défaillance de ce lien ou un des deux routeurs des extrémités provoque l'isolation complète entre ce sub-AS et la confédération. Prenons un autre exemple : si plusieurs sub-ASs sont connectés au sub-AS cœur par des sessions initiées exclusivement avec le même routeur, la défaillance de ce routeur va transformer la confédération en des îlots séparés. Dans les réseaux de l'épine dorsale, il y a une faible probabilité que deux composants tombent en panne en même temps, ou le deuxième tombe en panne avant que le premier soit réparé. Sous ces conditions, une topologie hub-and-spoke où chaque sub-AS non-transit a deux sessions indépendantes (pas corrélées par des supports physiques communs) empêche avec une grande probabilité l'isolement entre les sub-ASs. La topologie de la figure 1.7 satisfait cette condition et elle est donc caractérisée par une forte résilience.

Pour un réseau de topologie physique donnée, la conception de hub-and-spoke consiste à chercher un sub-AS central, puis assigner les autres sub-ASs. Les bords de chaque AS sont choisis en suivant la topologie physique pour tenir compte du bon fonctionnement de IBGP. Mais nous avons trouvé que cette méthodologie peut aboutir à un grand nombre de sessions EBGP en même temps qu'une autre conception peut servir des éléments physiques supportant ces sessions pour aider au bon fonctionnement de IBGP dans les sub-ASs. Le point faible de cette méthodologie est qu'elle met IBGP dans la deuxième place en donnant priorité à la consistance et la prédiction du routage. Pour une topologie physique non centralisée, la conception de hub-and-spoke perd son premier avantage et ignore une répartition plus conforme des sub-ASs. Plus le réseau est large et étendu, plus il est difficile de trouver une architecture hub-and-spoke convenable.

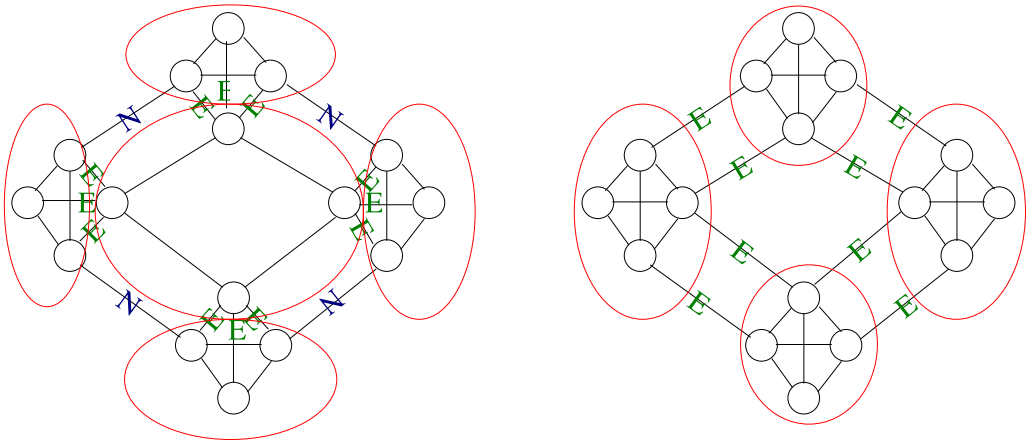

FIG. 1.8 – Comparaison entre deux architectures pour une topologie non centralisée

Pour la même topologie physique, nous exposons à la figure 1.8(a) l'architecture hub-and-spoke qui peut être employée. Remarquons le grand nombre de sessions EBGP (liens marqués par la lettre E) et que 4 liens physiques ne sont pas utilisés (liens marqués par la lettre N, les autres liens supportent IBGP). Remarquons aussi que chaque sub-AS

peut être isolé si le routeur de bord correspondant dans l'AS central tombe en panne. L'architecture à la figure 1.8(b) paraît plus élégante. Dans chaque sub-AS, les sessions IBGP possèdent des supports physiques indépendants. Chaque paire de sub-ASs sont liés par deux sessions EBGP et le routage intra-confédération est consistant.

1.5.3 Comparaison entre les réflecteurs de routeset les confédérations

Déterminer quelle conception utiliser, confédération ou réflexion des routes, n'est pas une décision simple. Bien que des organisations différentes ont expérimenté plusieurs niveaux de stabilité avec ces deux approches, la réflexion de routes paraît plus flexible à implanter et à maintenir et les fournisseurs ont gagné plus d'expérience à la traiter. Ainsi, elle est recommandée pour résoudre le problème de mise en échelle de IBGP. Cependant pour résoudre le problème d'instabilité de IGP dans les ASs à grande échelle, la confédération permet de déployer des IGPs indépendants dans les différents ASs et elle est la solution la plus favorable.

1.6 Conclusion

Le passage à l'échelle de IBGP pose un problème pour les grands domaines. La topologie de la maille complète n'est pas extensible. Les topologies alternatives sont les réflecteurs de route et la confédération. L'analyse de la fiabilité du travail de IBGP est difficile et la fiabilité est en relation avec la topologie logique employée. Normalement, remplacer la topologie de la maille complète conduit à une perte dans l'indépendance logique (c'est à dire plusieurs routeurs peuvent dépendre d'une seule session). La migration vers une des deux solutions est recommandée exclusivement pour les domaines de grande taille. Dans les petits domaines, les topologies alternatives peuvent avoir une plus grande résilience, ce qui est démontré dans [6] pour les réflecteurs de routes avec des topologies redondantes (réflecteurs redondants pour les mêmes clients et sessions redondantes entre clients). Nous pensons qu'IGP est fort dans ces domaines et il converge rapidement en réduisant la probabilité de l'échec d'une session pour une panne dans le chemin de son passage. Les petites différences dans le calcul de résilience seront insignifiantes et la re-configuration du réseau est toujours exclue.

Une fois la décision de migration vers les réflecteurs des routes est prise, les modèles des graphes et les métriques présentées dans ce chapitre sont à la base d'une méthode pour élire la topologie la plus fiable parmi toutes les topologies possibles (nous avons $\sum_{M=1}^{N}\sum_{i=0}^{M}(-1)^i\frac{(M-i)^N}{i!(M-i)!}$ variations pour grouper N nœuds en des clusters [11]). Le passage à l'échelle de IGP favorise fortement le choix de la confédération. Nous avons trouvé que les recommandations pratiques existantes pour optimiser la conception sont insuffisantes et qu'une méthodologie qui vise à employer une architecture de hub-and-spoke ne peut pas être appliquée dans le cas général. En effet, cette méthodologie propose un modèle de répartition des sub-ASs en ignorant l'optimisation des opérations IBGP à l'intérieur de chaque sub-AS. Des questions comme "combien de sub-ASs il faut employer ?" et "où sont les bords de chaque sub-AS ?" n'ont pas encore de réponses basées sur une

approche théorique. Dans le cadre de la résolution de cette problématique, viennent nos contributions présentées dans le chapitre suivant.

Chapitre 2

Contributions

2.1 Introduction

Étant donnée l'infrastructure physique d'un réseau souffrant d'un problème de passage à l'échelle, la conception d'une solution de réflexion de routes ou de confédération est la terminaison de deux étapes importantes :

1. élire la topologie logique la plus fiable par rapport aux éléments physiques existants ;
2. à un niveau plus haut, configurer les politiques BGP et profiter de ses compétences pour orienter les opérations du routage.

Après avoir étudié l'état de l'art, nous pouvons classer facilement les directives générales proposées dans la littérature sous l'une ou l'autre des deux étapes. Par exemple, nous pouvons utiliser la directive "garder les topologies physique et logique homologues le plus possible" dans la première étape, et des directives comme "fixer des métriques comparables (l'attribut MED) pour la sélection des routes", "modifier l'attribut NEXT-HOP avec attention" et "fixer des métriques IGP appropriées" avec la deuxième étape.

Notre sujet est en relation avec la première étape. Nous visons par l'optimisation de la topologie logique à former la base indispensable pour partir sur la deuxième étape. Dans ce chapitre, nous abordons l'approche théorique pour la conception d'une confédération. Nous exposons les différentes composantes de notre contribution : les modèles de graphes, la formalisation du problème et des contraintes, les métriques utilisées pour évaluer la fiabilité d'une topologie, une solution informatique et l'évaluation expérimentale correspondante. Notre travail est un complémentaire des approches théoriques étudiant la conception des réflecteurs de routes([11],[12]).

2.2 Modèles des graphes

Nous représentons le réseau physique dans un AS comme un graphe non orienté $G(V, E)$, où V représente l'ensemble des routeurs et E représente l'ensemble des liens physiques. Nous notons $(i, j) \in E$ l'arête entre le nœud $i \in V$ et $j \in V$. Généralement, il y a des routeurs qui n'exécutent pas BGP, nous représentons par V_p l'ensemble des routeurs qui sont des locuteurs BGP, $V_p \subseteq V$, et nous définissons $n = |V_p|$ comme le nombre

des locuteurs. Nous focalisons sur un domaine de transit où tous les routeurs exécutent BGP ($V = V_p$), et nous considérons que notre modèle peut être étendu simplement pour couvrir le cas général. Un modèle de fiabilité doit être en soi lié à la fiabilité des composants individuels c-à-d les routeurs et les liens physiques. La fiabilité d'un routeur décroît fortement si la consommation de ses ressources (que ce soit le CPU pour les traitement des routes, ou la mémoire pour maintenir les tables de routage) augmente au delà d'un seuil donné (en fonction des performances du routeur). La consommation des ressources est fonction du nombre de sessions manipulées simultanément par le routeur. Dans une confédération, excepté les routeurs du bord du domaine, chaque routeur doit gérer des sessions avec les locuteurs de son sub-AS au lieu de tous les locuteurs du domaine. Cette conception repose donc la contrainte déjà mentionnée et résout le problème de passage à l'échelle. Mais la défaillance temporaire d'un composant physique est une partie des opérations quotidiennes d'un réseau. Nous associons à chaque routeur i une valeur de fiabilité v_i définie comme la proportion du temps où i est en bonne santé. v_i peut être assignée grâce à l'histoire de surveillance ou estimée grâce aux performances du routeur. De même nous associons à chaque lien (i, j) une valeur de fiabilité w_{ij} définie comme la proportion du temps où (i, j) travaille proprement. S'il n'y a pas de lien entre deux routeurs i et j, alors nous prenons $w_{ij} = 0$.

La topologie logique est formée par k sub-ASs, chaque sub-AS est représenté par un sous-graphe et se voit assigner un numéro SAS, $1 \leq SAS \leq k$. Le modèle logique $G(V, E, f)$ est obtenu en caractérisant le modèle physique par une fonction $f : V \mapsto [1, k]$. f assigne à chaque nœud le sous-graphe qui le contient. La propriété principale de f est qu'elle divise le graphe en des sous-graphes connexes. En nous basant sur cette fonction, nous pouvons calculer le nombre des nœuds de chaque sous-graphe par la formule : $y(SAS) = card(\{i \in V; f(i) = SAS\})$. Le nombre des arêtes entre les nœuds du même sous-graphe peut être aussi calculé : $m(SAS) = card(\{(i, j) \in E; f(i) = f(j) = SAS\})$. Nous pouvons détecter si un routeur peut initier des sessions EBGP intra-confédération et il est parsuite un routeur de bord, par une fonction b : $b(i) = 1$ if $\exists j \in V; (i, j) \in E \wedge f(i) \neq f(j)$. Donc $b(i) = 1$ si i est un routeur de bord et $b(i) = 0$ autrement. Pour initier une session EBGP intra-confédération, deux routeurs de bords doivent résider dans deux sous-graphes différents. Nous utilisons une fonction s pour détecter cette propriété, $i, j \in V : s(i, j) = 1$ if $f(i) \neq f(j)$ et 0 autrement.

2.3 Énoncé du problème

Étant donné la topologie du réseau physique $G(V, E)$ d'un système autonome, nous devons trouver parmi toutes les topologies logiques de confédération celle qui assure la meilleure fiabilité.

i	A	G	B	F	D	C	E
f(i)	1	1	1	1	2	2	2
b(i)	0	0	1	1	0	1	1

TAB. 2.1 – La topologie logique comme solution

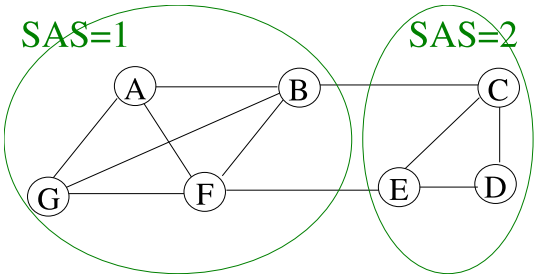

FIG. 2.1 – La topologie physique

Par exemple, nous donnons la topologie physique de la figure 2.1. Nous supposons qu'un ou plusieurs parmi les 7 routeurs n'ont pas les performances suffisantes pour manipuler 6 sessions concurremment. Nous voulons diviser Le AS en des sub-ASs en optimisant la fiabilité du protocole du routage. La solution pour cette topologie est représentée au tableau 2.1. Une justification théorique de ce choix ne peut pas être complète sans étudier les facteurs qui influencent sur la fiabilité de IBGP et de EBGP. Nous modélisons ces facteurs par une métrique appropriée accompagnée par trois contraintes essentielles.

2.4 La métrique de densité et les contraintes du problème

Nous définissons une métrique capable d'évaluer la différence entre une Clique [3] et un graphe de même taille (en terme du nombre de nœuds). La motivation de notre approche est qu'une Clique a la plus petite perte de sessions prévue (ESL) et la plus grande connectivité des arêtes (*edge connectivity*) parmi tous les graphes de même taille. La connectivité globale des arêtes pour un graphe non orienté est le nombre minimal des arêtes qu'il faut enlever pour couper le graphe en deux ou plus parties connexes [2]. Pour une Clique de n nœuds, la connectivité des arêtes est de $n-1$. Notre approche est de diviser le réseau en un petit nombre de sub-ASs denses. La notion de densité est utilisée en [9] pour caractériser la hiérarchie de l'Internet. Pour un graphe k-divisé (c'est à dire divisé en k sous-graphes connexes), nous définissons d'abord la densité d'un sous-graphe comme le rapport entre le nombre de ses arêtes et le nombre des arêtes nécessaire pour accomplir une Clique entre ses nœuds. Pour n nœuds, nous avons besoin de $\frac{n\times(n-1)}{2}$ arêtes pour accomplir une Clique.

$$D(SAS) = \frac{m(SAS)}{\frac{y(SAS)\times(y(SAS)-1)}{2}}$$

Nous définissons la densité du graphe comme la moyenne des densités de ses sous-graphes

$$D = \frac{\sum_{SAS=1}^{k} D(SAS)}{k}$$

Une topologie logique qui fait concentrer les arêtes dans les sous-graphes réduit en même temps le nombre des arêtes entre les sous-graphes et le nombre des sessions EBGP est minimisé.

Pour adresser la résilience de EBGP, nous introduisons ici la contrainte de *la fiabilité de la coupure*. Nous définissons la fiabilité de EBGP intra-confédération comme la somme des fiabilités des composants de l'infrastructure qui utilisent EBGP (routeurs de bord et lien physiques entre eux) et nous la notons R. R indique approximativement combien des composants utilisent EBGP et à quel degré ces composants sont fiables.

$$R = \sum_{i \in V} v_i \times b_i + \sum_{(i,j) \in E} w(ij) \times b(i) \times b(j) \times s(i,j).$$

Notre contrainte exige que R doit être plus grande qu'un certain seuil pesé par une fraction α à la fiabilité totale R_T. La fiabilité totale est définie comme la somme des fiabilités de tous les composants du réseau.

$$R_T = \sum_{i \in V} v_i + \sum_{(i,j) \in E} w_{ij}$$

La contrainte sera formulée sous la forme : $R > \alpha \times R_T$. Le but essentiel de cette contrainte est d'augmenter le nombre des éléments qui exécutent EBGP, et par suite les sessions EBGP et la redondance logique, ce qui diminue la probabilité d'isolation entre les routeurs (et ce qui est semblable à avoir des réflecteurs redondants pour le même cluster dans une topologie de réflexion). Les valeurs de fiabilité sont insérées au lieu du simple comptage pour des raisons d'évaluation plus que de conception. Par exemple, considérons les deux nœuds A et B de la figure 2.2 : dans 2.2(a), A et B sont connectés à travers une seule session, tandis que dans 2.2(b), les mises à jours de A peuvent arriver à B directement par la session A-B, ou indirectement par la session A-D et puis par IBGP. La redondance logique entre A et B (de même A et C ou A et D) est plus grande dans (b).

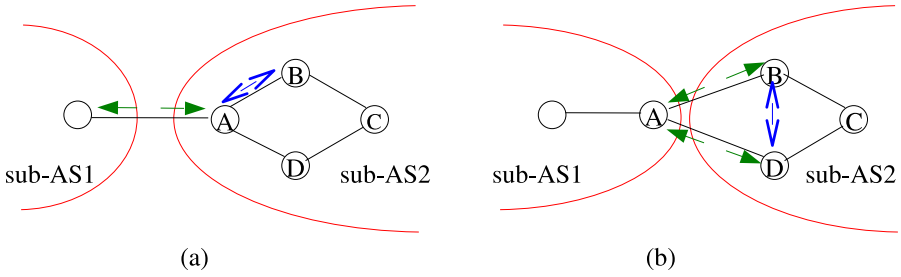

FIG. 2.2 – L'avantage d'augmenter le nombre des sessions intra confédération

La deuxième contrainte que nous utilisons consiste à limiter *le nombre de sub-ASs*. Le routage dans la confédération peut prendre des décisions sous optimales sans configurer manuellement les politiques de BGP. Quand le nombre de sub-ASs augmente, les avantages de IGP deviennent moins importants. Pour se convaincre de l'idée, il suffit d'imaginer le cas extrême où chaque routeur constitue un sub-AS et de se demander comment le routage peut être effectué sans la notion de la longueur des chemins. Ainsi nous choisissons de ne pas excéder un certain seuil pour le nombre de sub-ASs. Sans cette contrainte nous avons besoin d'un effort conséquent pour assurer la stabilité et l'efficacité sur le plan du routage.

Finalement, il est important de distribuer uniformément les routeurs sur les sub-ASs. De cette façon nous équilibrons le nombre de sessions que chaque routeur va manipuler, ce qui protège certains routeurs d'une consommation de ressources excessive, et nous équilibrons entre les IGP's qui travaillent dans les différents sub-ASs. La troisième contrainte est donc appelée la contrainte d'*équilibrage des charges*.

L'exemple de la figure 2.3 calcule la densité de 4 topologies logiques différentes pour la même topologie physique et montre comment elles peuvent être exclues à cause des contraintes.

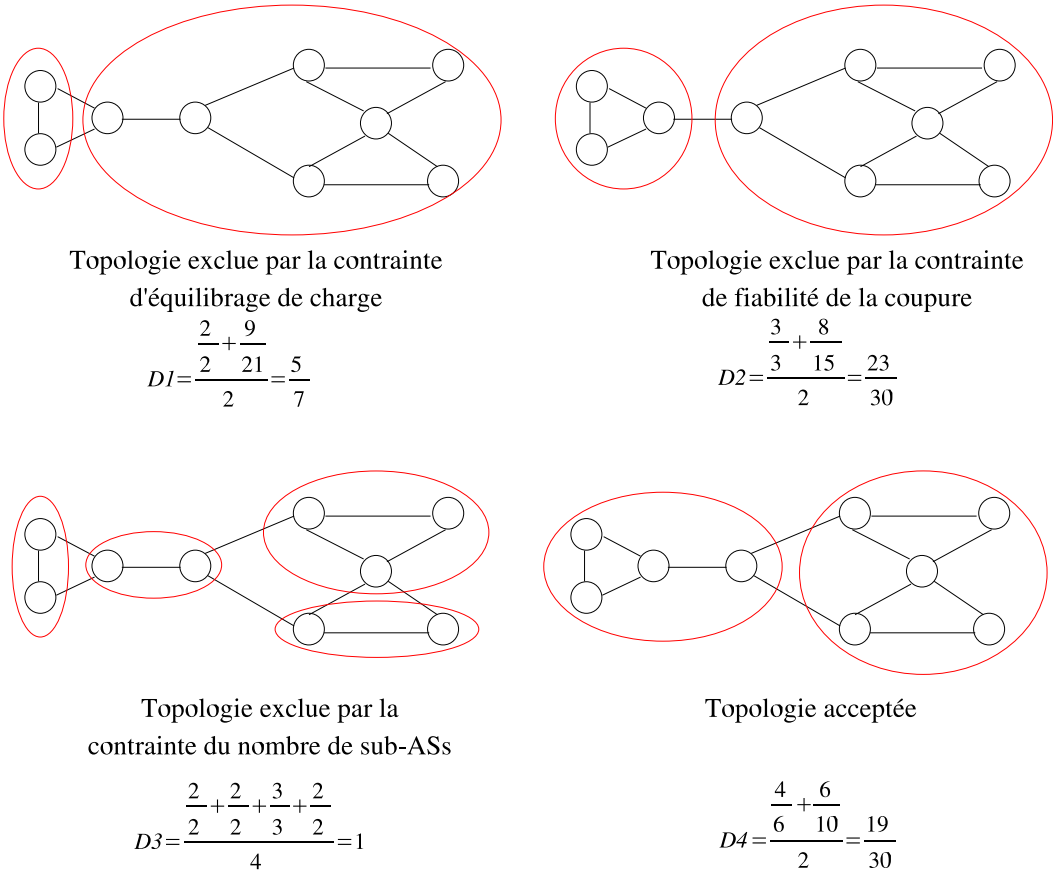

FIG. 2.3 – Application des contraintes et calcul des densités

2.5 Le problème de fiabilité de confédération-Densité (RC-D)

Étant donné un graphe non orienté $G(V, E)$, ainsi que les valeurs de fiabilité des routeurs et des liens v_i et w_{ij}, le problème RC-D (pour *reliable confederation-density*)

consite à trouver k et la k-coupure du graphe qui maximise la métrique de densité D tout en respectant les trois contraintes formulées comme suit :

1. la contrainte de fiabilité de la coupure : $R > \alpha \times R_T$ où α est choisi dans l'intervalle $[0, 0.1]$;
2. la contrainte de nombre de sub-ASs : $2 \leq k < \lceil \ln(n) \rceil$;
3. la contrainte d'équilibrage des charges : $\forall SAS; \beta \times \frac{n}{k} < y(SAS) < (2 - \beta) \times \frac{n}{k}$ où β est choisi de $[0.5, 0.9]$.

Nous pouvons choisir α et β et changer le seuil de k pour renforcer ou relâcher les contraintes. Un bon choix exige une expérience pratique et une étude des exemples historiques des confédérations.

2.6 Solution heuristique pour RC-D

Si k est fixe et le graphe va être divisé en k sous-graphes exactement, donc nous aurons le problème k-RC-D. Par la comparaison des solutions du problème k-RC-D pour k allant de 2 jusqu'à $\lceil \ln n \rceil$, et l'élection de la solution qui maximise notre métrique, nous résolvons le problème RC-D pour ce graphe. Notre solution est un algorithme randomisé du genre ([8]) qui utilise une technique de l'ensemble des solutions du problème de la K-Coupure minimale (*Min k-Cut*).

Notre solution HS fixe k d'abord et utilise une procédure randomisée appelée *Contract* en second lieu pour diviser le graphe en k sous-graphes connexes. La procédure *Contract* choisit d'une manière aléatoire une arête de E (la même probabilité pour toutes les arêtes). L'arête choisie est éliminée et les deux extrémités sont joints à un même méta-nœud. Les arêtes de chacun des deux extrémités appartiennent maintenant au nouveau méta-nœud. Cette contraction est répétée d'une manière itérative et s'arrête quand nous atteignons k méta-nœuds. Les nœuds compressés dans chaque méta-nœud sont retournés comme un sous-graphe connexe. La sortie de cette procédure est une topologie logique associant une fonction f au graphe. f est représentée par une liste qui assigne à chaque nœud dans V le SAS du sous-graphe qui le contient. Le pseudo code de *Contract* peut être simplifié comme suit :

Entrée : G(V,E)
TANT QUE $|V| > k$ *FAIRE*
    Choisir aléatoirement une arête (u, v) *de* E
    E=E-$\{(u, v)\}$
    SI (u est un méta-nœud) ET (v est un méta-nœud) ALORS
        Ajouter arêtes(v) à arêtes(u) (Toujours éliminer les arêtes de type (u, u)*)*
        Ajouter liste(v) à liste(u)
        V=V-$\{v\}$
    SINON
        SI (u est un méta-nœud) ALORS
            Ajouter arêtes(v) à arêtes(u)
            Ajouter v à liste(u)

V=V-{v}
SINON
SI (v est un méta-nœud) ALORS
Ajouter arêtes(u) à arêtes(v)
Ajouter (u) à la liste(v)
V=V-{u}
SINON
u = Nouveau méta-nœud
Ajouter arêtes(v) à arêtes(u)
Ajouter v à liste(u)
V=V-{v}
Sortie : k méta-nœuds chacun a une liste des nœuds y compressés.

Le travail de *Contract* est démontré sur une entrée simple dans la figure 2.4. Nous avons fixé k à 2 dans l'exemple, et le travail de *Contract* s'arrête quand V se réduit à un ensemble de 2 nœuds.

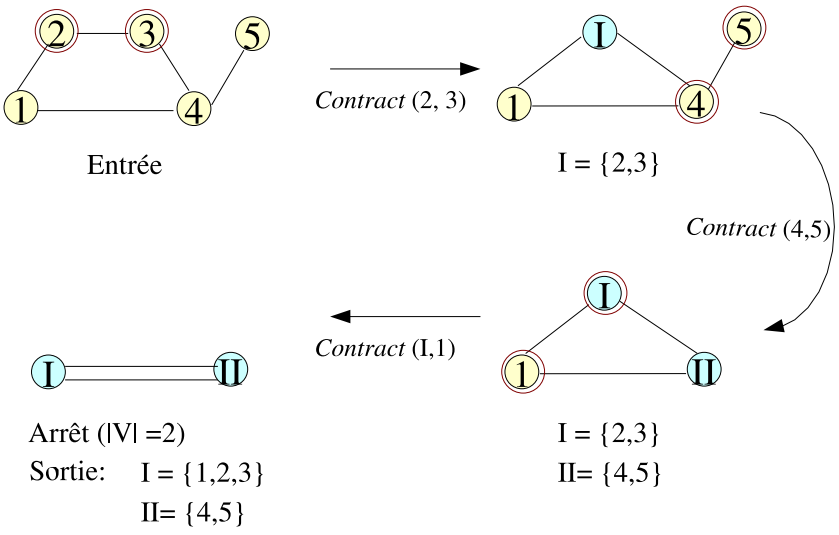

FIG. 2.4 – Le travail de *Contract* ($k = 2$)

A la suite, HS calcule la fiabilité de la coupure, et le nombre des routeurs dans chaque sous-AS $y(SAS)$. Si la topologie excède la contrainte de la fiabilité de la coupure ou la contrainte d'équilibrage des charges, HS lui assigne une densité nulle. Autrement, il calcule la densité de chaque sous-graphe et puis la densité moyenne. HS répète ce travail (*Contract* + calcul de la métrique) n^2 fois comme dans l'algorithme de k-coupure pour augmenter la chance d'être proche de la solution optimale. A la fin de cette boucle, HS sélectionne la densité maximale et la liste correspondante qui représente f comme la solution du problème k-RC-D.

Pour répondre au problème RC-D, HS assigne à k toutes les valeurs entières entre 2 et $\lceil \ln n \rceil$, résout pour chaque valeur le problème k-RC-D, et finalement retourne parmi toutes les solutions celle ayant la densité maximale. Ainsi, la complexité de notre solution

est $O(n^2(n^2+m)\ln n)$ parce que la complexité de la procédure *Contract* est $O(n^2)$ et le calcul de densité pour les différents sous-graphes de la topologie résultante se fait en $O(m)$, $m = |E|$ est le nombre des arêtes du graphe entier G. Le pseudo code de HS est le suivant :

POUR $k = 2$ *à* $\lceil \ln(n) \rceil$
    POUR topologie $= 1$ *à* n^2
        f[topologie]=Contract(G)
        SI (f[topologie] satisfait les contraintes) :
            D_top[topologie] = calcul_D(f[topologie])
        SINON
            D_top[topologie]=0
    D_k[k]= max(D_top)
D_opt=max(D_k)
RETOURNER(D_opt,k_opt,f_opt)

2.7 Résultats expérimentaux

Dans cette section, nous évaluons les résultats d'optimisation de l'algorithme HS par un test expérimental. Nous avons implanté un algorithme de force brutale (ET) qui prend la topologie physique et le nombre de sub-ASs k comme entrée, travaille en temps exponentiel, essaye toutes les combinaisons, produit toutes les topologies logiques possibles et retourne exactement la densité maximale possible comme sortie. La complexité de l'algorithme (ET) est k^n. Notre objectif est de comparer les résultats de HS avec ceux de ET.

Les topologies physiques des réseaux étaient produits à l'aide du générateur des topologies BRITE (*Boston university Representative Internet Topology gEnerator*) [7]. Nous avons utilisé BRITE parce qu'il est un des générateurs largement utilisés dans la communauté des recherches sur les réseaux et l'Internet. Nous avons choisi d'utiliser la distribution Heavy-tailed pour placer les nœuds et le modèle du Waxman pour les interconnecter. Ensuite, les fiabilités des liens physiques $\{w_{ij}\}$ étaient produites aléatoirement dans l'intervalle $[0, 0.9]$ et les fiabilités des routeurs $\{v_i\}$ étaient produits aléatoirement dans l'intervalle [0,0.99]. Pour les contraintes nous avons fixé α à $\frac{1}{n}$ et β à 0.5. Nous avons produit trente-trois topologies physiques : dix topologies pour chaque taille de dix, quinze et vingt nœuds et trois topologies pour la taille de vingt-cinq nœuds (pour cette taille, le temps d'exécution de ET pour une machine ordinaire commence à devenir très long(k^{25})). Pour chaque topologie, nous avons décidé de couper le graphe en deux sous-graphes (nous avons fixé k à 2), et nous avons exécuté les deux algorithmes (HS et ET). Il est plus dur à ET de travailler avec $k = 3$ pour les topologies de vingt nœuds ou plus. Pour faire le test avec $k = 3$, nous avons utilisé seulement les vingt premières topologies de tailles dix et quinze nœuds. Les deux diagrammes de la figure 2.5 montrent la différence des réponses entre les deux algorithmes.

Pour une topologie d'entrée, la densité de la topologie logique de sortie générée par l'algorithme ET est notée par D_{ET} et la densité de celle générée par l'algorithme HS est

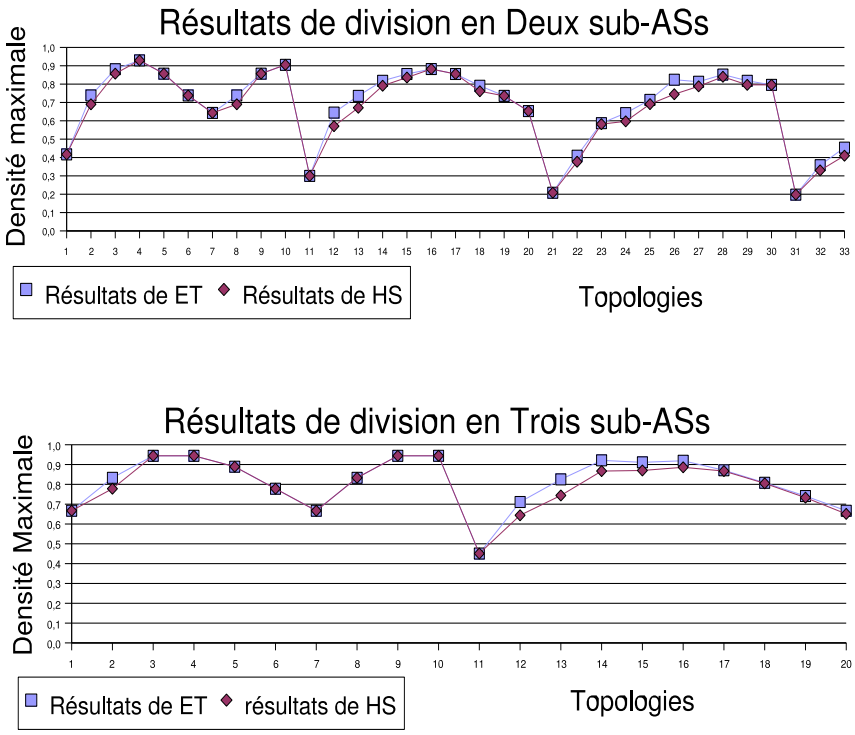

FIG. 2.5 – Résultats expérimentaux[1]

notée par D_{HS}, l'erreur relative peut être donc calculée par la formule :

$$e_r = \frac{D_{ET} - D_{HS}}{D_{ET}} \times 100$$

Nous avons utilisé l'erreur relative pour son potentiel à comparer les erreurs commises sur deux topologies d'entrée de concentration en arêtes différente.

Pour chaque ensemble de topologies qui ont la même taille, nous avons comparé l'erreur relative moyenne et l'erreur relative maximale. Les résultats sont donnés dans les deux diagrammes de la figure 2.6. Après interpréter les diagrammes, nous avons trouvé que les résultats de HS ne sont pas très loin de ceux de ET. L'erreur relative moyenne ne dépasse pas 6% et l'erreur relative maximale ne dépasse pas 12% pour la division en deux sub-ASs, et respectivement 4% et 10% pour la division en trois sub-ASs. Nous avons conclu que l'algorithme HS peut être une solution acceptable pour approcher le problème RC-D, soit en terme de complexité ou en terme d'exactitude, sachant que nous pouvons trouver des solutions plus élaborées dans le futur tout en se basant sur l'état de l'art de la résolution de problème de k-coupure.

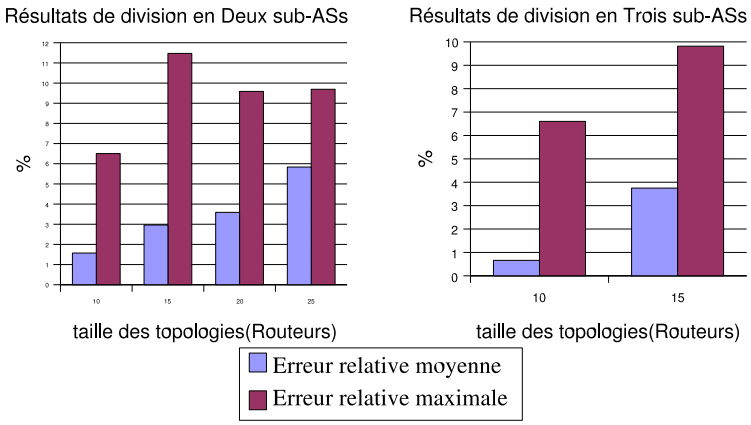

FIG. 2.6 – Résultats expérimentaux[2]

2.8 Une métrique alternative : connectivité de Steiner relative

La métrique de densité donne une indication sur le nombre des arêtes utilisées dans chaque sub-AS, sans savoir la forme de répartition de ces arêtes. D'autre part, notre approche ne prend pas en compte les routeurs du bord du domaine entier (en liaison avec autres ASs). Le nombre de chemins indépendants entre deux routeurs du bord d'un même sub-AS est d'une importance supérieure au nombre de chemins indépendants entre deux autres routeurs, parce que les routeurs de bord jouent un rôle important dans le transit du trafic et des informations du routage à l'intérieur et à l'extérieur de la confédération. Avoir plusieurs chemins indépendants entre deux routeurs de bord peut augmenter d'une manière considérable la capacité de TCP à surmonter (par retransmission et re-routage) une panne physique dans le chemin de la session qui les inter-connecte. Pour prendre en compte ce facteur, nous proposons une métrique alternative qui différencie entre deux sous graphes de même densité et qui est maximale quand la densité est maximale (cas d'une Clique).

La connectivité globale des arêtes pour un graphe non orienté est définie comme le nombre minimal des arêtes dont l'effacement divise le graphe en des parties connexes [4]. La connectivité des arêtes de plusieurs nœuds spécifiés dans le graphe peut être considérablement différente de la connectivité globale du graphe. Par exemple, la connectivité entre deux nœuds peut être C tandis que la connectivité globale du graphe est 1. Pour un graphe non orienté et un sous ensemble spécifié des nœuds (appelés *terminals*), la connectivité Steiner est le nombre minimal des arêtes dont la destruction découpe le graphe au moins en deux composants connexes d'une telle manière que les terminals sont répartis sur deux ou plusieurs composants [2]. Pour une Clique de taille n, la connectivité Steiner est égale à $n-1$ quel que soit les terminals spécifiés.

Nous appelons notre propre métrique la connectivité de Steiner relative(CSR), et nous la définissons pour un sous-graphe(SAS) de taille ($y(SAS)$) et un sous-ensemble

spécifié(B) de ses nœuds comme le rapport entre la connectivité Steiner(CS) de ce sous-graphe divisé par la connectivité Steiner pour une Clique de même taille :

$$CSR(SAS, B) = \frac{CS(SAS, B)}{y(SAS) - 1}$$

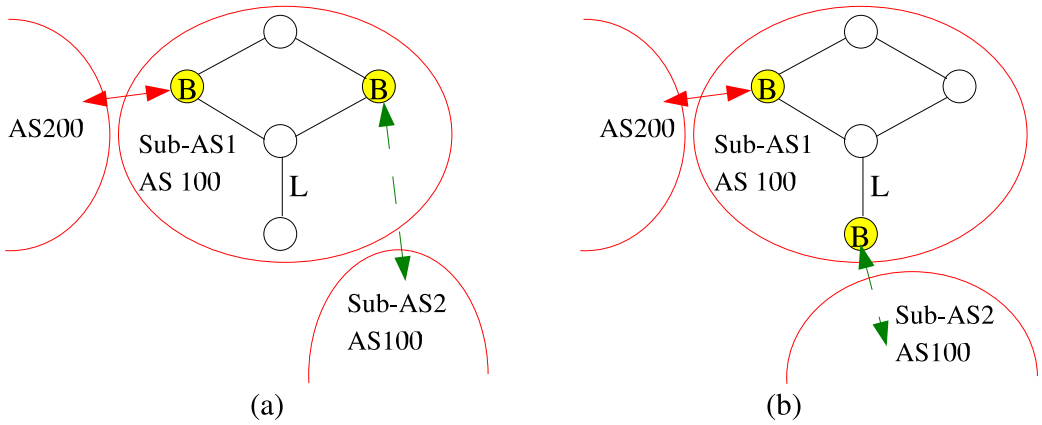

FIG. 2.7 – Différenciation entre deux sub-ASs de même densité

Notre approche est de calculer la connectivité Steiner relative des arêtes au lieu de la densité pour chaque sous-graphe en spécifiant les routeurs de bord de niveau AS ou de niveau sub-AS comme l'ensemble des terminals. Un exemple très simple est illustré dans la figure 2.7. Les nœuds intitulés par la lettre B sont les terminals. Le sub-AS1 de (a) est plus fiable que celui de (b). En effet, supposons que la session L tombe en panne dans les deux réseaux. Dans le réseau (a), un routeur serait isolé du réseau et il n'y a pas un grand problème. Cependant dans le réseau (b), le AS200 et le sub-AS2 de AS100 seraient isolés. La redondance physique pour les deux routeurs de bord dans (a) peut protéger (avec une certaine probabilité) la session relativement importante. Cela est traduit directement par l'effet que la connectivité Steiner relative de (a) est $2/4$ et elle est supérieure à la connectivité Steiner de (b) qui est $1/4$. Cependant la densité de sub-AS1 est égale à $5/10$ dans les deux réseaux et IBGP est supporté au même degré. Remarquons que la connectivité globale est égale à 1 dans les deux réseaux.

Un algorithme rapide pour calculer la connectivité Steiner est publié dans [2]. Cet algorithme est basé sur une construction efficace des emballages des arbres qui généralise le théorème d'Edmonds et sa complexité est $O(C^3 n \log n + m)$ pour un graphe de n nœuds et m arêtes, C étant la connectivité Steiner pour l'ensemble des terminals spécifiés. En remplaçant le calcul de la densité qui se fait en $O(m)$ par le calcul de la connectivité Steiner pour un sous-graphe, notre algorithme HS augmente de complexité. Nous n'avons pas implanté cet algorithme et interprété ses résultats à cause de la durée limitée de notre stage.

Même si l'emploi de la connectivité Steiner relative à la place de la densité dans HS pourrait ne pas donner des résultats de grande différence. La connectivité Steiner reste un outil important pour l'évaluation de la compétence d'un domaine ou d'un sous-domaine

à supporter les opérations du transit, ou pour localiser où l'insertion des arêtes peut augmenter la connectivité entre les routeurs de bord d'un domaine ou d'un sous domaine.

Conclusion générale

Le protocole BGP est aujourd'hui le protocole responsable du routage dans l'Internet. Le fonctionnement de ce protocole est essentiellement dirigé par sa configuration. Cette configuration a un double rôle. Elle établit une topologie virtuelle (logique) (*overlay network*) qui dirige le processus d'échange des routes et elle permet l'implantation des politiques de routage par les opérateurs. L'évaluation et la conception des topologies virtuelles de IBGP étaient les motifs de notre travail.

Nous avons commencé ce manuscrit par une brève description de la structuration de l'Internet et de BGP. Dans la partie état de l'art, nous avons expliqué les facteurs qui agissent sur la fiabilité de IBGP. Nous avons introduit les définitions et les techniques des solutions alternatives, les avantages et les inconvénients de chacune, et les critères qui favorisent le choix de l'une sur l'autre. Pour les réflecteurs des routes, nous avons résumé les approches théoriques existantes pour l'évaluation des topologies et les méthodologies de conception basées sur l'optimisation des métriques appropriées. Pour les confédérations, nous avons exposé les recommandations pratiques de la conception et nous avons demontré l'insuffisance de ces recommandations et l'absence d'une approche spécialisée sur le plan théorique.

Dans la partie contributions nous avons étudié la particularité des conditions de fiabilité des confédérations. Nous avons proposé les modèles des graphes qui peuvent representer l'entrée (topologie physique) et la sortie (topologie logique) de notre problème. Nous avons posé le problème d'optimisation d'une manière formelle en associant une métrique quantitative. Nous avons formulé les contraintes indispensables pour la solution de notre problème. Ensuite, nous avons proposé une solution informatique pour résoudre le problème dans un temps linéaire. Nous avons implanté notre solution et nous avons validé le degré de son exactitude par expérience sur des topologies générées avec des tailles différentes. A la fin de cette approche, nous avons proposé une métrique alternative qui évalue la connectivité entre les routeurs de bord.

Notre travail est complémentaire des travaux existants sur les réflecteurs de route. Il peut être d'une grande utilité avec le besoin progressif pour le déploiement des confédérations dans les domaines de routage de grande échelle. Il peut être poursuivi dans le même sens que l'optimisation de la topologie logique de IBGP en fonction de la répartition et les caractéristiques des éléments physiques de l'AS. Il peut également être poursuivi par l'étude de la conception des politiques BGP en profitant de ses grandes capacités pour orienter les opérations de routage vers un régime stable et fiable.

Bibliographie

[1] T. Bates, R. Chandra, and E. Chen. Bgp route reflection - an alternative to full mesh ibgp. RFC 2796, Internet Engineering Task Force, ,United States, April 2000. http ://www.rfc-archive.org.

[2] Richard Cole and Ramesh Hariharan. A fast algorithm for computing steiner edge connectivity. In *STOC '03 : Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 167–176, New York, NY, USA, 2003. ACM Press.

[3] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2001.

[4] Harold N. Gabow. A matroid approach to finding edge connectivity and packing arborescences. In *STOC '91 : Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 112–122, New York, NY, USA, 1991. ACM Press.

[5] S. Halabi and McPherson D. *The definitive BGP resource : Internet Routing Architectures Second Edition*. Cisco Press, 2000.

[6] L. Xiao L. and K. Nahrstedt. Reliability models and evaluation of internal bgp networks. In *IEEE INFOCOM 2004, Hong Kong, China*, March 2004.

[7] A. Medina, A. Lakhina, I. Matta, and J. Byers. Brite : Universal topology generation from a user's perspective (user manual). Technical Report BU-CS-TR-2001-003, Boston University, 2001. "http ://www.cs.bu.edu/brite/".

[8] Rajeev Motwani and Prabhakax Raghavan. Randomized algorithms. *SIGACT News*, 26(3), 1995.

[9] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz. Characterizing the internet hierarchy from multiple vantage points. Technical Report CSD-01-1151, University of California at Berkeley, Berkeley, CA, USA, 2001.

[10] P. Traina, D. McPherson, and J. Scudder. Autonomous system confederations for bgp. RFC 3065, Internet Engineering Task Force, ,United States, February 2001. "http ://www.rfc-archive.org".

[11] L. Xiao, J. Wang, and K. Nahrstedt. Optimizing ibgp route reflection network. In *IEEE International Conference on Communications (ICC 2003), Anchorage, Alaska*, May 2003.

[12] L. Xiao, J. Wang, and K. Nahrstedt. Reliability-aware ibgp route reflection topology design. Technical Report UIUCDCS-R-2003-2375/UILU-ENG-2003-1762, Department of Computer Science University of Illinois at Urbana-Champaign, August 2003.

[13] R. Zhang and M. Bartell. *BGP Design and Implementation : Practical guidelines for designing and deploying a scalable BGP routing architecture.* Cisco Press, 2004.

Annexe

L'article exposé dans cet annexe a été soumis au 5ème atelier international IEEE sur l'exploitation et la gestion de IP (IPOM2005) à la date du 6/Mai/2005 et nous attendons le résultat de l'évaluation à la date du 3/Juillet/2005.