Thèse de Doctorat de l'université Paris VI Pierre et Marie Curie

Spécialité

SYSTÈMES INFORMATIQUES

présentée par

Cheikh Ahmadou Bamba GUEYE

pour obtenir le grade de

Docteur de l'université Pierre et Marie Curie

Localisation géographique des hôtes dans l'Internet basée sur la multilatération

Soutenue le 08 Décembre 2006 devant le jury composé de

M. :	Pierre	SENS	Prof. à l'université Pierre et Marie Curie	Président
MM. :	Walid	DABBOUS	Directeur de recherche INRIA	Rapporteurs
	Philippe	OWEZARSKI	HDR au Laboratoire LAAS-CNRS	
MM. :	Mark	CROVELLA	Prof. à Boston University	Examinateurs
	Steve	UHLIG	HDR à Delf University	
	Artur	ZIVIANI	Chercheur au LNCC Brazil	
M. :	Serge	FDIDA	Prof. à l'université Pierre et Marie Curie	Encadrant

Thèse de Doctorat de l'université Paris VI Pierre et Marie Curie

Spécialité

SYSTÈMES INFORMATIQUES

présentée par

Cheikh Ahmadou Bamba GUEYE

pour obtenir le grade de

Docteur de l'université Pierre et Marie Curie

Localisation géographique des hôtes dans l'Internet basée sur la multilatération

Soutenue le 08 Décembre 2006 devant le jury composé de

M. :	Pierre	SENS	Prof. à l'université Pierre et Marie Curie	Président
MM. :	Walid	DABBOUS	Directeur de recherche INRIA	Rapporteurs
	Philippe	OWEZARSKI	HDR au Laboratoire LAAS-CNRS	
MM. :	Mark	CROVELLA	Prof. à Boston University	Examinateurs
	Steve	UHLIG	HDR à Delf University	
	Artur	ZIVIANI	Chercheur au LNCC Brazil	
M. :	Serge	FDIDA	Prof. à l'université Pierre et Marie Curie	Encadrant

A mes parents

et

ma femme Seynabou.

Une civilisation qui s'avère incapable de résoudre les problèmes
que suscite son fonctionnement est une civilisation décadente.
Une civilisation qui choisit de fermer les yeux à ses problèmes
cruciaux est une civilsation atteinte.
Une civilisation qui ruse avec ses principes est une civilisation moribonde.
Aimé CÉSAIRE

Remerciements

Je tiens spécialement à remercier :

Serge FDIDA, de m'avoir permis, de venir faire le DEA Réseaux, puis une thèse. Il a toujours su allier rigueur et bonhomie. J'oublierai jamais ses remises en cause.

Artur ZIVIANI, pour son soutien depuis mon stage de DEA, le temps et l'intérêt qu'il a consacrait à nos nombreuses discussions. Ce fut un plaisir de travailler avec lui.

Steve UHLIG, de m'avoir permis de travailler avec lui, aussi, d'avoir accepter de participer à mon jury de thèse. Je désire témoigner à Plop toute ma gratitude. Je me rappellerai toujours de nos longues discussions existentielles.

Mark CROVELLA, pour l'honneur qu'il me fait d'être dans mon jury, pour l'intérêt qu'il porte à ces travaux, et pour m'avoir permis de travailler avec lui.

Walid DABBOUS et Philippe OWESARSKI, qui ont accepté d'être rapporteurs sur ma thèse.

Pierre SENS, d'avoir accepté de prendre part à ce jury.

Tous les membres de l'équipe NPA. Particulièrement Farid et Benoit qui ont fait du bureau C681 un havre de quiétude. Nos diverses discussions me manqueront.

Badou, Doumbia, Papis, Boly, M. Niang, M. et Mme Salvat pour toutes ces années passées ensemble.

Mes parents, mes sœurs et mes frères qui m'ont accompagné dans toutes les entreprises de ma vie. Je remercie enfin ma femme Seynabou qui par sa présence et son amour m'a encouragé tout au long de ces dernières années.

Résumé

L'inférence de la localisation géographique d'hôtes dans l'Internet à partir uniquement de leur identifiant a vu son importance économique grandir ces dernières années. Elle permet l'émergence de nouvelles applications variées basées sur la localisation. La localisation physique d'un hôte se fait soit grâce à des mesures de délai, soit via l'association d'une localisation physique à un hôte ou un bloc d'hôtes. Dans cette thèse nous abordons le problème de la localisation basée sur des mesures de délai. Ces mesures de délai sont faites par des serveurs sondes vers un hôte cible et vers des hôtes références, qui sont des hôtes dont on connaît la localisation géographique. Jusqu'à présent, l'estimation de la localisation d'un hôte cible était fournie par la position géographique de l'hôte référence qui lui est le plus proche. Ainsi, le nombre d'endroits possibles où l'on peut localiser un hôte cible est égal au nombre d'hôtes références, conduisant à un espace discret de réponses.

Dans cette thèse, nous proposons une nouvelle technique, CBG (Constraint-Based Geolocation), basée sur la multilatération. La multilatération permet d'obtenir un espace continu d'endroits possibles où l'on peut localiser un hôte cible. Pour ce faire, CBG transforme les mesures de délai en distance géographique estimée, et attribue une zone de confiance à chaque hôte localisé. Cela permet aux applications, qui utilisent CBG, d'évaluer la fiabilité de l'estimation par rapport à leurs exigences.

Nous montrons également l'impact du délai de *buffering* (temps de traitement des paquets dans les routeurs) sur les mesures de délai et par conséquent sur la géolocalisation. En se basant sur l'outil *traceroute*, les résultats obtenus montrent qu'en tenant compte du buffering, avant de transformer les mesures de délai en distances géographiques estimées, nous améliorons la précision de l'estimation de localisation ainsi que la zone de confiance associée à chaque estimation.

Pour réduire le nombre d'hôtes références nécessaire pour localiser un hôte cible, nous proposons une technique hybride qui utilise une base de données, contenant des préfixes d'adresses IP et leur localisation géographique, et des mesures de délai.

Mots-clés

géolocalisation, multilatération, estimation de délai, traceroute, Internet

Abstract

Geolocation of Internet hosts enables a diverse and interesting new class of location-aware applications. This thesis focuses on geographic location of Internet host based on delay measurements. Previous measurement-based approaches use reference hosts, called landmarks, with a well-known geographic location to provide the location estimation of a target host. This leads to a discrete space of answers, limiting the number of possible location estimates to the number of adopted landmarks.

In contrast, we propose Constraint-Based Geolocation (CBG), which infers the geographic location of Internet hosts using multilateration with distance constraints to establish a continuous space of answers instead of a discrete one. However, to use multilateration in the Internet, the geographic distances from the landmarks to the target host have to be estimated based on delay measurements between these hosts. CBG accurately transforms delay measurements to geographic distance constraints, and then uses multilateration to infer the geolocation of the target host. In contrast to previous approaches, our method is able to assign a confidence region to each given location estimate. This allows a location-aware application to assess whether the location estimate is sufficiently accurate for its needs.

Currently, measurement-based geolocation techniques disregard the buffering delays that may be introduced at each hop along the path taken by probe packets. Relying on traceroute measurements, we show that leveraging buffering delay estimation improves accuracy in the measurement-based geolocation of Internet hosts as well as the confidence that the geolocation service associates to each estimation.

We propose an hybrid approach in order to reduce the chosen set of landmarks to locate a given Internet host. This hybrid approach uses a database, that contains IP prefix addresses and its own geographic location, and delay measurements.

Keywords

geolocation, multilateration, delay measurement, traceroute, Internet

Table des matières

Chapitre 1

Introduction

Internet est un vaste réseau assez complexe permettant l'interconnexion d'hôtes de part le monde. Sa croissance durant cette dernière décennie à fait naître de nouveaux besoins. D'après une étude faite par *Comscore* [1] (entreprise spécialisée dans l'analyse d'audience) au mois de mars 2006, on estimait à 694 millions le nombre d'internautes. Cela représente 14 % de la population mondiale âgée de 15 ans et plus. Avec l'augmentation des moteurs de recherche et de distributeurs de contenu, de nouveaux services dits de "proximité" ont fait leur apparition. Ces services sont basés sur la localisation géographique des clients uniquement à partir de leur adresse IP (*géolocalisation*).

Les moteurs de recherche tels que *Google* [2], *Yahoo !* [3], et *MSN* [4], les trois poids lourds qui se partagent 80% du marché mondial de la recherche sur Internet, fournissent chaque jour des milliards de réponses aux questions de leurs utilisateurs. Rien qu'aux États Unis, *Google US* a reçu 1120 requêtes par seconde durant le mois de juin 2006 [1]. Les moteurs de recherche, associant une géolocalisation dans leur service (*Google Maps*, *MSN Virtual Earth*), deviennent de plus en plus populaires auprès des annonceurs qui peuvent cibler plus efficacement leur publicité. Les réponses fournies par l'outil de recherche peuvent différer en fonction du critère géographique, afin d'augmenter leur pertinence. Par exemple, une requête de livraison de repas peut entraîner une réponse différente de l'outil de recherche si l'internaute est à Paris ou à Dakar...

Le marketing sur les moteurs de recherche devient un gigantesque marché : il devrait rapporter 1,8 milliard d'euros en Europe en 2006 et atteindre 2,2 milliards d'euros dès 2008, selon les prévisions de *Forrester* [5]. Par exemple, les milliards engrangés par ces nouveaux rois de l'information pourraient représenter un tiers des investissements publicitaires en 2010 en Europe. Les publicités sont tellement bien ciblées, que les millions d'internautes les tolèrent et même les apprécient.

Ces publicités ciblées sont basées sur le contenu, mais aussi principalement sur la localisation géographique des internautes.

Les techniques qui permettent de localiser les hôtes dans l'Internet à partir de leur adresse IP peuvent utiliser deux types de mesure :

1. **Les mesures passives :** "aucune" mesure n'est générée dans le réseau. Les techniques basées sur les mesures passives peuvent être qualifiées de non intrusives. Dans le cadre de la géolocalisation, ces techniques utilisent des bases de données qui contiennent des blocs d'adresses IP et leur information de localisation. Le serveur de localisation envoie une requête à la base de données lorsqu'un hôte cible désire se faire localiser.
2. **Les mesures actives :** elles consistent à envoyer des paquets sondes dans le réseau pour déterminer par exemple la qualité de service (délai, RTT, pertes, débit), la topologie interne d'un réseau, la tomographie, les points de congestion. Parmi les outils de mesures actives nous pouvons citer l'outil *ping*, l'outil *traceroute*, l'outil *MGEN*. Dans le cadre de la localisation géographique des hôtes dans l'Internet, les outils ping et traceroute sont utilisés.

La connaissance de la distribution géographique des nœuds constituant Internet permettra aussi de quantifier et de mesurer les propriétés géographiques qui sont derrière cette structure complexe.

1.1 Motivations et problématique

La possibilité de connaître la localisation physique d'un hôte à partir uniquement de son identifiant (adresse IP) ouvre la voie à des services inédits [6, 7, 8, 9, 10, 11, 12, 13, 14]. La localisation est devenue une information à grande valeur ajoutée, que cela soit d'un point de vue économique ou militaire. De nombreux services dépendent et se servent de la position des utilisateurs pour rendre des services personnalisés. Dans l'Internet, la correspondance géographique des adresses IP est nécessaire aux services tels que :

- **La publicité ciblée sur les pages WEB :** les internautes sont désormais exigeants envers les publicités avec l'abus de fenêtres pop-up et d'animations que l'on note actuellement. Toute publicité liée à la zone géographique de l'internaute est immédiatement pertinente pour lui. Ainsi, les annonceurs peuvent définir plusieurs stratégies de marketing. Utilisée intelligemment, la publicité ciblée peut aider un site WEB à augmenter considérablement son taux de clics et, par conséquent, ses revenus.
- **L'utilisation de "redirects" :** un redirect permet d'orienter automatiquement un internaute vers un site national en fonction de sa localisation. Un internaute se connectant à partir de la France sur le site portail d'une

multinationale peut se retrouver dirigé d'office vers le site national en *.fr*. La pratique du redirect est par exemple utilisée de manière intensive par *Google* [2]. Le redirect forcé est également intéressant pour respecter les politiques marketing différenciées (prix, produit,...) en fonction des pays ou zones géographiques.

- **La diffusion restreinte de contenu :** supposons qu'un grand événement sportif doive être diffusé en direct sur Internet. Si nous avons une politique de diffusion, une application utilisant un service de localisation géographique peut déterminer quels sont les clients qui ont le droit d'accéder au contenu diffusé suivant leur emplacement. Toutefois un organisme de régulation est nécessaire. Ainsi cet organisme peut envisager que l'évènement sportif soit diffusé partout ailleurs excepté l'endroit où il se déroule. De même, on peut stipuler qu'un contenu sera mis seulement à la disposition d'internautes provenant de certaines régions ou bien situés dans le même endroit que le serveur de contenu.
- **Identification basée sur la localisation du client :** avec la croissance du commerce en ligne, de nombreux cas de fraudes sont notées actuellement. Un système de vérification de la localisation géographique du client avant l'acceptation de toute transaction commerciale peut être établi. Cela permettra de contrecarrer l'utilisation frauduleuse de numéro de cartes de crédit. Ses endroits préétablis, à partir desquels on peut accepter des transactions, peuvent être établis pour un client donné. Un client peut émettre le souhait d'accepter une transaction commerciale, faite à partir de sa carte de crédit, qu'à partir de sa ville de résidence. S'il arrive que le client doit voyager, il peut avertir le gestionnaire de la carte de crédit pour qu'il configure le système de vérification géographique suivant sa nouvelle destination. Ainsi, une transaction établie à partir d'un endroit quelconque peut entraîner un refus pour non respect des clauses.
- **Lutte contre la cyber-criminalité :** il ne semble pas nécessaire d'insister ici sur l'enjeu que représente la lutte contre la *pédocriminalité* sur Internet. On estime que le nombre de sites à caractère pédophile a augmenté de 70% en un an dans le monde soit plus de 13000 sites pédophiles dans le monde [15]. La généralisation de l'accès à l'Internet, toutefois, facilitant les communications et les échanges, et aussi en offrant un certain sentiment d'anonymat, joue un rôle essentiel dans le phénomène. Un service de localisation géographique peut permettre à identifier les individus qui téléchargent et/ou proposent des contenus ou bien ceux qui ont un comportement suspect lors de discussion en ligne.

La localisation peut s'appliquer aussi dans le cadre des réseaux *ad hoc* sans fils. Elle est nécessaire, dans le cas où les décisions de routage sont basées sur la position des nœuds. En effet, un routage est dit géographique lorsque les décisions

de routage sont basées sur la position des nœuds. Dans un réseau ad hoc, pour faire du routage géographique, un nœud source connaît toujours la position du nœud destinataire. Pour se faire, soit les nœuds possèdent un moyen de localisation natif comme le système de positionnement par satellites (GPS) [16], soit un service de localisation doit être utilisée ou bien un système logiciel comme le protocole *SimPA* (*Simple Positioning Algorithm*) [17]. Chaque nœud du réseau va obtenir sa position physique par rapport aux autres nœuds grace à ce service de localisation géographique. Puisque tous les nœuds connaissent leur position entre eux, un nœud source est capable d'envoyer un paquet à tout autre nœud du réseau en le transmettant à un nœud qui est physiquement proche du destinataire. Parmi tous les nœuds qui reçoivent ce message, seul celui ou ceux qui sont le plus près géographiquement de la cible le retransmettront. Cette forme de routage basée sur la position des nœuds est appelé routage géographique [18, 19]. La possibilité de faire parvenir des informations à une destination sans connaître la route au préalable est un avantage pour ce type de routage.

Ces applications peuvent avoir différentes exigences par rapport à la précision de l'estimation de localisation.

Étant donné un identifiant associé à un hôte Internet peut-on déterminer sa position géographique ? Les identifiants (adresses IP) utilisés pour identifier les hôtes terminaux dans l'Internet sont alloués de manière arbitraire et il n'existe pas de relation entre une adresse IP et la position géographique [20] de l'équipement qui possède cette adresse. De plus la structure topologique du réseau d'un FAI (Fournisseur d'Accès Internet) fournit peu d'informations sur sa couverture géographique. Dès lors, inférer la localisation géographique d'un hôte dans l'Internet est un véritable défi. En outre, les techniques de localisation géographique, basées sur des mesures de délai, se basent sur une possible corrélation entre délai et distance géographique. Bien que cette corrélation soit assez robuste [21, 22], le délai mesuré entre une source et une destination peut souffrir de différentes sources de distorsions. Ces sources de distorsions (violation de la règle de l'inégalité triangulaire [23], présence de goulots d'étranglement, non linéarité des chemins entre les hôtes [24],...) peuvent ajouter un délai additionnel, dans les mesures de délai, qui peut fausser la corrélation entre délai et distance géographique. Ainsi, les techniques de géolocalisation basées sur les mesures de délai doivent nécessairement tenir compte de ces sources de distorsions pour fournir une bonne estimation de localisation. Le trafic engendré par les mesures de délai dans le réseau (passage à l'échelle), ainsi que le temps de réponse nécessaire pour la localisation d'un hôte cible, sont aussi des facteurs qu'il faut prendre en compte.

1.2 Contributions

Nous proposons d'appliquer la *multilatération* dans l'Internet [25] pour inférer la position géographique des hôtes cibles. En effet, la multilatération permet d'estimer une position en utilisant un nombre suffisant de distances à partir de quelques points immobiles. Dès lors, elle fournit un ensemble continu d'endroits où on peut localiser la cible au lieu d'un espace discret de réponses comme les techniques précédentes de géolocalisation. Toutefois, pour pouvoir appliquer la multilatération dans l'Internet, il faut transformer les mesures de délai, entre les hôtes références (hôte dont on connaît la position géographique) et les hôtes cibles, en distances géographiques estimées. Pour transformer les mesures de délai, la technique proposée, *CBG* ("*Constraint-Based Geolocation*") [26, 27], implémente un algorithme d'auto-calibration pour tenir compte des possibles distorsions qui peuvent s'ajouter aux mesures de délai. Cette auto-calibration permet de capturer la meilleure relation pouvant exister entre délai et distance géographique dans le réseau. Ensuite, CBG applique la multilatération pour inférer la zone géographique dans laquelle l'hôte cible se trouve. L'approximation de cette zone géographique par une heuristique -utilisation d'un polygone - permet d'obtenir une surface appelée zone de confiance. Ainsi, CBG est capable de fournir une zone de confiance à chaque estimation de localisation d'un hôte cible. Cela permet aux applications, qui utilisent CBG, d'évaluer la fiabilité de l'estimation par rapport à leurs exigences. Ainsi, chaque application peut définir le niveau de précision qu'elle désire.

Nous montrons également l'impact du délai de "*buffering*" (temps de traitement des paquets dans les routeurs) sur les mesures de délai et par conséquent la géolocalisation. La technique proposée dans [28], *GeoBuD*, tient compte de la topologie du réseau avant de transformer les mesures de délai en distances géographiques estimées. Ainsi, GeoBuD tient compte du délai de buffering de chaque routeur intermédiaire sur le chemin entre les hôtes références et l'hôte cible. La technique GeoBuD, en se basant sur l'outil traceroute, estime et supprime ce délai de buffering au niveau du délai avant de faire la transformation de ce délai en distance géographique. Les résultats montrent qu'en tenant compte du délai de buffering, nous obtenons une réduction de la zone de confiance associée à chaque estimation de localisation, et par conséquent une meilleure estimation de la localisation des hôtes cibles.

Les applications dans l'Internet, ayant besoin d'un service de géolocalisation, exigent un temps de réponse assez rapide. Cette réponse se mesure par rapport au temps de chargement d'une page WEB (environ 1 à 3 secondes en moyenne). Pour pallier cette exigence, nous avons mis en place une technique hybride qui utilise une base de données et des mesures de délai. La base de données contient des préfixes d'adresses IP et leur information de localisation géographique. Cette

technique hybride, grace à une heuristique que nous avons développé, choisie les hôtes références les plus proches géographiquement de la cible. Elle permet ainsi d'obtenir un temps de réponse assez rapide et de limiter aussi le trafic généré dans le réseau par les mesures de délai faites par les hôtes références vers les hôtes cibles. En outre, un ensemble de 20 hôtes références est suffisant pour avoir une bonne estimation de localisation.

1.3 Sommaire

La thèse est organisée comme suit : le chapitre 2 passe en revue les différents domaines dans lesquels un service de localisation géographique peut être nécessaire. Nous présentons également les techniques d'estimation de localisation existantes dans le domaine des réseaux ad hoc et de l'Internet. Dans le domaine de l'Internet, nous proposons d'utiliser la multilatération dans le chapitre 3, pour inférer la position des hôtes dans l'Internet uniquement à partir de leur adresse IP. Pour pouvoir appliquer la multilatération dans l'Internet, et inférer la position des hôtes cibles, il faut transformer les mesures de délai faites par les hôtes références (hôte dont on connaît la position géographique) vers les hôtes cibles en distances géographiques. La technique *CBG*, que nous proposons, utilise la multilatération qui permet d'avoir un espace continu d'endroits où on peut localiser les hôtes cibles. Dans le chapitre 4, nous décrivons l'heuristique utilisée par l'approche CBG pour inférer la position des hôtes cibles. Nous évaluons aussi la technique CBG. Nous montrons que la technique CBG dépasse en précision les précédentes techniques de géolocalisation basées sur les mesures de délai. Le chapitre 5 examine l'impact du temps de traitement, au niveau des routeurs (*buffering*), sur les mesures de délai et par conséquent sur la géolocalisation. Nous envisageons d'utiliser l'outil *traceroute* pour estimer le délai de buffering sur le chemin entre les hôtes références et l'hôte cible. Toutefois, l'utilisation des mesures de délai génèrent un trafic non négligeable dans le réseau. Ainsi, dans le chapitre 6 nous proposons une technique hybride qui va associer une technique de mesures passives (utilisation de base de données) et une technique de mesures actives (mesures de délai) pour inférer la position des hôtes cibles. Enfin, le chapitre 7 conclut cette thèse.

Chapitre 2

État de l'art

Avec l'augmentation des moteurs de recherches et des distributeurs de contenus, des nouveaux services ont fait leur apparition, notamment des services dits de "proximité" basés sur la localisation géographique des clients. A partir d'un identifiant du client, nom ou adresse IP, ces services tentent de déterminer sa position physique. Ces services de localisation géographique peuvent avoir leur application dans différents domaines tels que les réseaux ad hoc ou Internet.

Dans ce chapitre, nous montrons dans la section 2.1 la différence entre un système de positionnement et un service de localisation. Dans la section 2.2 nous présentons un état de l'art des différentes techniques de localisation existantes dans les domaines tels que les réseaux cellulaires, les réseaux ad hoc, et Internet.

2.1 Différence entre système de positionnement et service de localisation

Un système de positionnement permet à un utilisateur ou à un équipement de connaître sa position. Le système de positionnement le plus connu reste de loin le GPS. Toutefois pour contrer l'hégémonie du GPS américain, l'Union Européenne, la Chine et la Russie ont mis sur pied leur propre système de positionnement Galiléo [29], GLONASS [30], et Beidou [31] respectivement. Pour inférer la position d'un nœud, des satellites en orbite émettent des signaux radio en direction de la terre. Pour estimer sa position, un nœud doit posséder un équipement permettant la reconnaissance de ces signaux. La position est ensuite estimée par trilatération (mesure de distance). Ces systèmes de positionnement par satellites, bien que assez performant en milieu ouvert, restent inefficaces en mileu fermé. Cette inefficacité est due à la perturbation du signal radio. La visibilité des satellites

n'est pas toujours possible au fond d'une vallée ou dans un bâtiment. Puisque les équipements mobiles, les équipements embarqués, et les réseaux sans fil sont de plus en plus utilisés , des systèmes de positionnement fonctionnant en milieu intérieur - dans un bâtiment par exemple - ont été proposés. Nous pouvons citer par exemple *Active Badge* [32], *Crickets* [33] et RADAR [34].

Un service de localisation est une extension du système de positionnement. En effet, un service de localisation permet à un nœud de connaître la position de tout noeud appartenant au système tandis qu'un système de positionnement permet à un utilisateur ou un équipement de connaître uniquement sa propre position. Par exemple, dans un réseau ad hoc où un routage géographique y est appliqué, un nœud source doit connaître la position géographique de n'importe quel nœud destinataire pour être capable d'étiqueter les paquets avec la position de la destination. Dans cette thèse, nous nous sommes focalisés sur les services de localisation liés à Internet.

2.2 Domaines d'application de la localisation

Nous passons en revue dans cette section les différentes techniques de localisation qui permettent de localiser des nœuds ou des utilisateurs dans le domaine du GSM, des réseaux ad hoc, et de l'Internet.

2.2.1 GSM

Un système de réseaux cellulaire comme le *GSM* (*Global System for Mobile Communication*) [35] utilise généralement deux bases de données pour gérer la mobilité des utilisateurs : le *HLR* (*Home Location Register*), qui tient à jour les données de l'abonné (par exemple position de l'abonné dans le réseau), et le *VLR* (*Visitor Location Register*), qui gère le client dans la cellule où celui-ci se trouve. Chaque utilisateur se voit alloué un numéro unique d'identité internationale, l'*IMSI* (*International Mobile Subscriber Identity*), utilisé par le réseau pour la transmission des données. Pour chaque utilisateur qu'il gère, le HLR possède son IMSI. Il connaît également le VLR dont dépend le mobile à un instant donné. Le VLR enregistre les informations de localisation des mobiles. Ainsi, la base de données VLR contient des données dynamiques qui lui sont transmises par le HLR avec lequel elle communique lorsqu'un abonné entre dans la zone de couverture du centre de commutation mobile auquel elle est rattachée. Lorsque l'utilisateur quitte cette zone de couverture, ses données sont transmises à un autre VLR ; sa position dans le réseau est ainsi connue à chaque instant donné.

2.2.2 Les réseaux Ad Hoc

Un réseau mobile ad hoc, appelé généralement MANET (*Mobile Ad hoc NETwork*), consiste en une grande population, relativement dense, d'unités mobiles qui se déplacent dans un territoire et dont le seul moyen de communication est l'utilisation des interfaces sans fils, sans l'aide d'une infrastructure préexistante ou d'administration centralisée. La topologie du réseau peut changer à tout moment. Elle est donc dynamique du fait de la mobilité prononcée des terminaux. Un des soucis majeurs des protocoles de routage classiques [36, 37, 38, 39] est la découverte des coordonnées des destinations, qui en général génère une surcharge de trafic assez importante. Ainsi, l'introduction d'une nouvelle métrique - la localisation - peut être assez bénéfique. Un routage est dit géographique lorsque les décisions de routage sont basées sur la position des nœuds. Parmi les propositions existantes, on distingue deux catégories de protocoles de routage : celle utilisant l'information de localisation afin d'améliorer des protocoles déjà existants (on peut désigner ces protocoles par "Aide au routage") et les protocoles de routage géographique.

2.2.2.1 Les aides au routage

Ce type de protocole ajoute des fonctions supplémentaires aux protocoles existants. Ces améliorations portent sur le nombre de messages de découverte de routes envoyés. La géolocalisation permet de délimiter un périmètre de recherche dans lequel le protocole de découverte de routes sera plus efficace. Par exemple on peut citer le protocole *LAR* (*Location Aided Routing*) [40] basé sur *DSR* (*Dynamic Source Routing*) [37] qui est un protocole réactif basé sur le routage à la source (*source routing*).

Avec le protocole LAR, lorsqu'une route, entre une source S et une destination D (Fig. 2.1), n'est plus valide du fait de la mobilité des nœuds, il est possible de faire une découverte locale des routes quand l'information de localisation est disponible. Dans la zone de rupture de la précédente route, LAR route les paquets vers les nœuds du réseau les plus proches de la destination, tout en étant dans la zone de couverture du nœud précédent. Une fois la nouvelle route établie, on utilise à nouveau un protocole de routage comme DSR. Le protocole LAR permet ainsi un routage géographique et optimise le nombre de messages émis lors des échecs de routes.

2.2.2.2 Les protocoles de routage géographique

Il existe deux types de routage géographique qui sont :

1. Le routage géographique hiérarchique : le *GGAR* (*GeoCast - Geographic Addressing and Routing*) [41] est une réflexion basée sur une étude indiquant

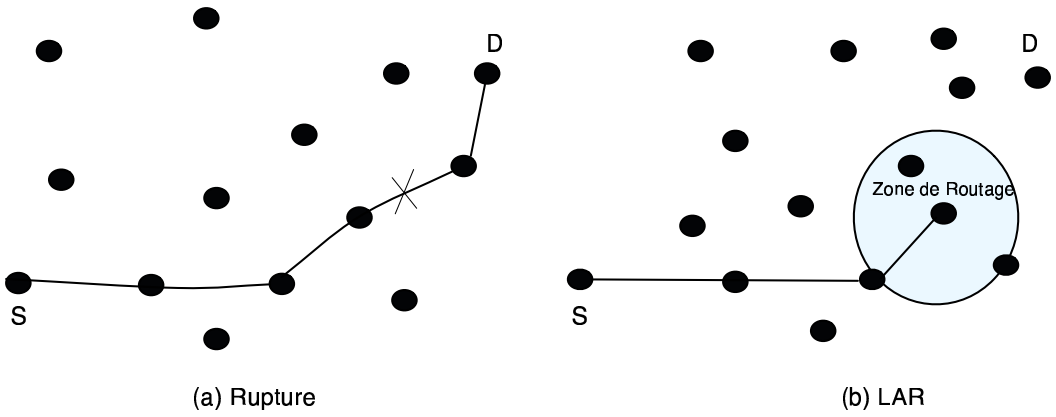


FIG. 2.1 – Application du LAR.

une possible utilisation des informations de géolocalisation en complément de l'adressage IP déjà existant [42]. Le routage dans un tel type de réseau est effectué de manière hiérarchique (Fig. 2.2) et est décomposé en trois domaines distincts :

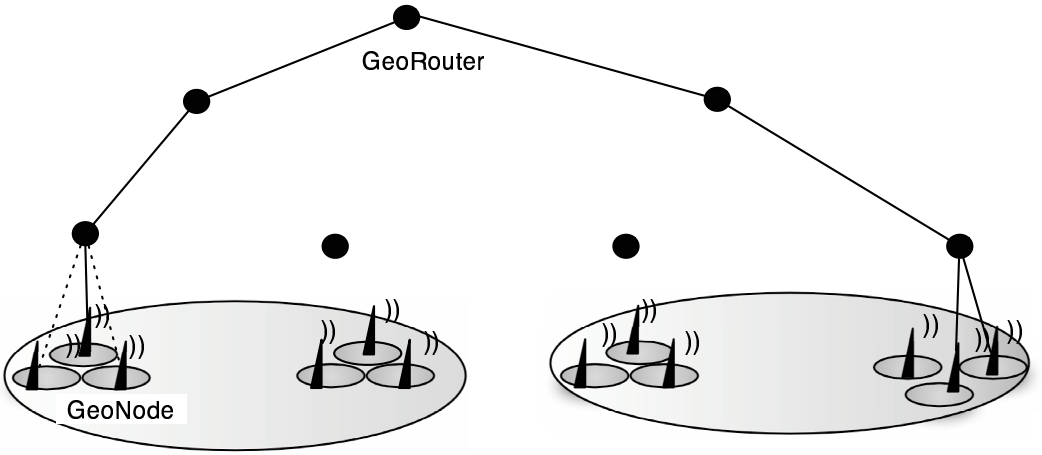


FIG. 2.2 – Exemple de routage géographique hiérarchique.

Les GeoRouters : ces routeurs prennent en charge le routage entre les zones géographiques. Les GeoRouters sont capables de détecter automatiquement les différents interfaces réseaux, le type d'interface (filaire ou sans fil), les autres GeoRouters, et les programmes des GeoNodes. Les paquets sont routés selon leur information de localisation.

Les GeoNodes : ces nœuds sont les points d'accès du réseau. Leur principale fonction est de faire du "*store and foward*". Par ce mécanisme,

les paquets peuvent être retransmis périodiquement, et ainsi toujours atteindre le destinataire. On évite ainsi les demandes de retransmission intempestives de paquets par le noeud source.

Les GeoHosts : Ils sont en fait des démons résidant dans le terminal de l'utilisateur. C'est uniquement par leur intermédiaire que la réception et l'envoi de messages géographiques sont possibles. Ils fournissent au client la possibilité d'utiliser le routage géographique lorsque celui-ci est disponible. Les GeoHosts notifient également à l'utilisateur l'adresse du GeoNode auquel il appartient.

2. Le routage géographique simple : le *GPSR* (*Greedy Perimeter Stateless Routing for Wireless Networks*) [43] et le *GRP* (*Geographical Routing*) [18] utilisent un routage géographique non hiérarchique. Ces deux protocoles proposent une même approche globale de l'utilisation des informations de localisation. Nous avons aussi des protocoles géographiques, qui assurent la délivrance des paquets à la destination [44], basés sur les graphes de Gabriel [45].

2.2.3 Internet

Dans l'Internet, l'estimation de la localisation géographique des hôtes est un véritable défi. Il n'existe aucune relation entre l'identifiant d'un hôte, nom ou adresse IP, et sa position géographique. La *RFC* (*Request For Comments*) 1876 [46] propose d'ajouter des informations de localisation dans les noms *DNS* (*Domaine Name Server*). Cependant cette proposition ne fut pas largement adoptée, car c'était un fardeau pour les administrateurs d'ajouter des enregistrements de localisation dans les bases de données DNS. Par ailleurs, la vérification de la véracité des informations enregistrées n'est pas une chose facile.

L'adoption de la distance géographique entre les hôtes Internet comme métrique de distance est évalué dans [47]. Lakhina et al. [21] ont dégagé des bases pour les futurs générateurs de topologie Internet grâce à l'étude de la distribution géographique des équipements (routeurs, liens, AS) dans l'Internet.

Dans ce qui suit, nous passons en revue les différentes techniques d'estimation de localisation géographique existantes.

2.2.3.1 Utilisation de bases de données

C'est à travers des bases de données *Whois* que des outils comme *IP2LL* [13] et *NetGeo* [48] tentent d'inférer la localisation géographique d'un hôte à partir de son adresse IP. En effet, avec les bases de données Whois on peut obtenir des informations administratives et techniques sur les noms de domaine, les adresses IP, et les politiques de routage des opérateurs. Pour la localisation géographique

des adresses IP, on peut utiliser la base de données des 5 RIR (*Regional Internet Registries*) existants (RIPE NCC, ARIN, APNIC, LACNIC, AfricNIC). Les RIRs assurent la coordination administrative et technique des fonctions adressage et routage à travers le réseau Internet (attribution d'adresses IP, de numéros AS). Le format des données utilisé dans les bases de données Whois est un format libre. Ainsi, nous avons deux formatages : le format de type RIPE NCC d'un coté (attribut :valeur), le format *Network Solutions* de l'autre (user friendly, programmer enemy). Une même requête sur des serveurs différents à toutes les chances de donner un résultat différent. En outre, les informations contenues dans les bases de données Whois peuvent être désuètes et inexactes.

Si une organisation possède un bloc d'adresses IP assez large et géographiquement dispersé, les bases de données Whois peuvent contenir une seule entrée pour l'ensemble de ce bloc [49]. Par exemple, le bloc d'adresses 4.0.0.0/8 est alloué à l'organisation *Genuity* et toute requête faite par une adresse IP appartenant à ce bloc au niveau de la base de données Whois d'ARIN, donne comme localisation la ville de Broomfield (CO) aux États Unis. Ainsi, une requête de localisation adressée à la base de données fournie comme réponse la position géographique de l'entité administrative qui gère ce bloc d'adresses IP, bien que les hôtes cibles peuvent être localisés de manière dispersée à l'intérieur de ce bloc. Cette requête peut se faire soit en interrogeant le port 43 du serveur Whois en TCP (*Transmission Control Protocol*), soit via une interface WEB (certains ne fournissent que ce type d'accès).

Nous avons aussi des services de localisation tels que *EdgeEscape* d'*Akamai* [50] et *TraceWare* de *Digital Island* [51] qui se basent sur une tabulation entre la taille des APs et leur localisation géographique correspondante enregistrée dans une base de données. Toutefois, les techniques qui se basent sur la tabulation sont difficiles à gérer et à mettre à jour.

2.2.3.2 Techniques basées sur les noms DNS

Ces techniques infèrent la localisation d'un hôte à partir des noms fournis par le DNS de l'hôte cible ou des routeurs qui lui sont proches. Parmi ces techniques basées sur les noms DNS, nous pouvons citer *GeoTrack* [52], *VisualRoute* [53], *GTrace* [54], et le projet *SarangWorld* [55]. Les noms DNS dans l'Internet contiennent parfois certaines indications sur la localisation. Ces informations de localisation peuvent avoir des granularités différentes. Par exemple à l'échelle d'une ville (`bcr1-so-2-0-0.Paris.cw.net` indique la ville de Paris située en France), d'un état (`www.state.ca.us` indique l'état de Californie aux États Unis), ou d'un pays (`www.ucad.sn` indique le Sénégal). Toutefois, la plupart des FAI n'utilisent pas un nommage standard. Ce nommage reste à l'appréciation de l'administrateur. Les auteurs de la technique GeoTrack dans [52] notent pour la

seule ville de Chicago (IL) au moins 12 différents codes qui peuvent la représenter (chcg, chcgil, cgcil, chi, chicago,...). Les routeurs peuvent avoir aussi le code d'un aéroport (sjc2-cw-oc3.**sjc**.above.net) , ou d'un pays (asd-nr16.**nl**.kpnqwest.net) contenu dans leur nom DNS. Certains FAI, utilisent assez souvent le code de l'aéroport de la ville dans lequel ils se trouvent comme convention de nommage. Ces techniques basées sur les noms DNS utilisent l'outil *traceroute* [56]. Le traceroute permet de déterminer les nœuds intermédiaires sur le chemin entre un serveur sonde et l'hôte qu'on veut localiser. Par exemple un traceroute exécuté à partir du LIP6 (localisé à Paris) vers l'UCL (localisé à Louvain-la-neuve) donne :

1 132.227.74.33 (132.227.74.33) 0.199 ms 0.273 ms 0.256 ms
2 r-olympe.lip6.fr (132.227.109.254) 1.238 ms 1.481 ms 1.198 ms
3 r-jusren.reseau.jussieu.fr (134.157.254.126) 0.938 ms 1.083 ms 1.453 ms
4 gw-rap.rap.prd.fr (195.221.127.181) 1.606 ms 2.085 ms 15.340 ms
5 jussieu-g0-1-165.cssi.renater.fr (193.51.181.102) 1.216 ms 1.563 ms 1.115 ms
6 nri-c-pos2-0.cssi.renater.fr (193.51.180.158) 1.993 ms 2.936 ms 1.842 ms
7 nri-b-g14-0-0-101.cssi.renater.fr (193.51.187.18) 2.355 ms 6.136 ms 2.263 ms
8 renater.rt1.par.fr.geant2.net (62.40.124.69) 2.540 ms 2.123 ms 2.170 ms
9 so-5-0-0.rt1.lon.uk.geant2.net (62.40.112.106) 29.880 ms 9.453 ms 10.876 ms
10 so-2-0-0.rt1.ams.nl.geant2.net (62.40.112.137) 17.780 ms 18.200 ms 17.482 ms
11 belnet-gw.rt1.ams.nl.geant2.net (62.40.124.162) 21.291 ms 20.914 ms 20.806 ms
12 oc192.m160.core.science.belnet.net (193.191.1.1) 24.020 ms 21.849 ms 77.339 ms
13 oc48.m20.access.lln.belnet.net (193.191.1.198) 21.403 ms 21.832 ms 21.670 ms
14 ucl-1.customer.lln.belnet.net (193.191.11.10) 21.786 ms 21.403 ms 21.562 ms
15 planetlab2.info.ucl.ac.be (130.104.72.201) 26.176 ms 21.417 ms 21.348 ms

Ces techniques extraient l'information de localisation contenue, si elle existe, au niveau des noms DNS des routeurs intermédiaires sur le chemin. Pour cela, elles font une recherche d'alignement de séquence de caractères ("*string matching*") sur les noms DNS des routeurs. Toutefois, faire un string matching sans tenir compte au niveau du nom DNS la position de la chaine à rechercher serait inapproprié. Par exemple, le code *charlotte* correspond à la ville de Charlotte (NC) dans l'est des États Unis. Faire un string matching sur le nom DNS *charlotte.ucsd.edu*, sans tenir compte de la position de la séquence de caractères à chercher, peut indiquer comme localisation de ce nœud la ville de Charlotte,

alors qu'il se trouve à San Diego (CA) dans l'ouest des États Unis. A travers ces observations, la technique GeoTrack [52] implémente une règle de string matching pour chaque FAI, qui spécifie la position à laquelle l'information de localisation doit apparaître, si elle existe, au niveau des noms des routeurs associés à ce FAI. Ainsi, chaque nom de routeur est divisé en un ou plusieurs séquences séparées par des points. La règle de string matching correspondant à chaque FAI spécifie à quelle position du nom du routeur, il faut chercher l'information de localisation correspondante. Par exemple, la règle du FAI *Sprintlink* spécifie que si le code de l'information de localisation existe , il est toujours placée à la première séquence en partant de la gauche du nom du routeur. Par exemple le nom du nœud *sl-bb10-sea-9-0.sprintlink.net* contient le code *sea* qui indique Seattle. Toutefois, différents FAI, assez souvent, ne respectent pas leurs conventions de nommage. Ainsi, la recherche d'alignement de séquence de caractères devient très difficile à faire ou bien conduit à des résultats erronés. En outre, la technique GeoTrack ne possède pas de règles pour tous les FAI existants.

Un routeur est dit "reconnaissable" si sa localisation géographique peut être déduite à partir de son nom DNS. Le dernier routeur reconnaissable sur le chemin entre le serveur sonde et l'hôte cible, donne sa position géographique comme estimation de localisation de l'hôte cible. L'inférence de la localisation géographique à partir des noms DNS présente certains inconvénients. En effet, tous les noms DNS ne contiennent pas d'information de localisation, limitant ainsi le nombre de routeurs reconnaissables. Il n'y pas de standardisation concernant le nommage des routeurs et les FAI utilisent leur propre convention. Ceci rend difficile l'extraction des informations de localisation. L'estimation de localisation peut être imprécise car le dernier routeur reconnaissable qui donne sa position comme estimation n'est pas forcément proche de la cible. L'autre problème est le fait que certains hôtes se connectent via un *proxy* ou un *firewall*. Si l'on fait un traceroute, c'est au niveau du proxy ou du firewall qu'il s'arrêtera. Ainsi, si l'hôte est assez éloigné par rapport au proxy ou au firewall, il y a une imprécision dans l'estimation de localisation, car c'est celle du firewall ou du proxy que l'on obtient.

2.2.3.3 Technique basée sur le clustering

La technique *GeoCluster* [52] proposé par Padmanabhan et Subramanian se base sur la notion de cluster [57] qui définit un groupe de clients proche topologiquement et sous l'autorité d'une même administration de contrôle. GeoCluster se base sur l'hypothèse que tous les hôtes qui se trouvent à l'intérieur d'un même cluster sont co-localisés, *i.e.*, les hôtes forment un *cluster géographique*. Ainsi, connaissant la localisation géographique de quelques hôtes à l'intérieur d'un cluster, GeoCluster infère la localisation géographique du cluster tout entier.

2.2.3.3.1 Identification des clusters géographiques Les clients d'un même cluster sont identifiés grace aux informations des AP (adresses préfixes) et *netmask* extraites des tables de routage *BGP* ("*Border Gateway Protocol*") [58, 59, 60]. En effet, les hôtes dans l'Internet sont regroupés par groupe et chaque groupe est sous l'autorité d'un AS. Un AS peut correspondre à un campus universitaire, une entreprise ou un FAI. Les informations de routage que fournissent un AS sont agrégées. Par exemple, les routes vers les hôtes d'un domaine administratif peuvent être annoncées aux autres hôtes d'Internet en préfixe d'adresses agrégé comme 132.227.0.0/16, plutôt d'annoncer les 65536 adresses IP individuelles que cet AP contient. Il faut noter qu'un AS peut annoncer un ou plusieurs APs pour des raisons politiques et/ou de répartition de charge.

Les APs obtenus à partir des tables de routage BGP permettent d'identifier des clusters topologiques comme montré dans [57]. GeoCluster assimile un AS à un AP. Puisqu' un AS correspond à une localisation géographique, alors à un AP on peut correspondre une zone géographique. Pour identifier le cluster géographique auquel appartient l'hôte cible, GeoCluster détermine d'abord son cluster topologique. Elle utilise les adresses préfixes contenues dans les tables de routage BGP et une base de données qui contient des informations (adresse IP,localisation géographique). Ainsi, dans la base de données, chaque adresse IP est associée à sa position géographique sous forme de couple. La détermination d'un cluster géographique se fait en trois étapes :

- Extraire l'adresse IP de l'hôte.
- Effectuer le plus long assortiment de préfixe des tables BGP en utilisant la base de données IP-Localisation.
- Classer le client dans le même cluster que les adresses IP dont leurs adresses préfixes ressemblent le plus à la sienne.

Ainsi, si la localisation géographique de quelques hôtes dans le même cluster géographique est connue, alors cette localisation est utilisée comme celle de l'hôte cible. Si l'AS correspond à un FAI qui couvre une large zone géographique, les APs annoncées via BGP sont plus spécifiques car pouvant correspondre à ses propres clients. Certains gros FAI (*AT&T*, *Sprint*, *UUNet*, etc.) n'annoncent que des APs fortement agrégées pour des raisons de passage à l'échelle. Dans ce cas, un AP pourrait correspondre à une très large zone géographique et comme GeoCluster assimile les APs comme cluster géographique, une mauvaise localisation pourrait s'en suivre. Ainsi, en considérant le routage inter-domaine obtenu à partir de BGP, une autre approche se basant sur un nouveau algorithme appelé *Subclustering* a été développée dans [52] pour solutionner ce problème. Cette variante de GeoCluster qui incorpore cet algorithme est appelée (BGP+*subclustering*).

2.2.3.3.2 Énoncé de l'algorithme Sub-clustering La variante de la technique GeoCluster, BGP+subclustering, dépend seulement du routage inter-domai-

ne BGP. Toutefois, la nouvelle idée introduite dans cette méthode est de diviser les APs qui ont une large portée géographique en utilisant les informations IP-Localisation se trouvant dans sa base de données. Pour chaque AP obtenue à partir de *eBGP* ("*Exterior Border Gateway Protocol*"), on va consulter la liste des associations IP-Localisation dont les adresses IP dépendent de cette AP. Si un consensus se dégage, on déclare l'AP comme étant un cluster géographique. Par contre si on n'observe pas de consensus on subdivise l'AP en deux. Par exemple l'AP 152.153.0.0/16 sera subdivisé en AP 152.153.0.0/17 et 152.153.128.0/17 et le test de consensus est répété pour chaque AP, ainsi de suite jusqu'à ce qu'un consensus se dégage. A la fin, on obtient une association entre les APs (d'origine et celles issues des partitions) et leur localisation.

L'efficacité de GeoCluster dépend de la véracité et de la représentativité des informations se trouvant dans la base de données IP-Localisation. Ces informations fournies par les utilisateurs peuvent être désuètes ou peu fiables.

2.2.3.4 Techniques basées sur les mesures de délai

Nous avons des techniques qui exploitent une possible corrélation entre délai et distance géographique pour inférer la localisation géographique des hôtes dans l'Internet. Par exemple, la technique *GeoPing* [52] se base sur ce concept. Nous avons aussi des techniques d'estimation de distance qui permettent d'estimer la distance entre deux hôtes par des mesures de délai. Toutefois, dans cette section nous nous focalisons beaucoup plus sur les techniques de localisation géographique basées sur les mesures de délai (par exemple GeoPing) plutôt que sur les techniques d'estimations de distance.

2.2.3.4.1 Techniques d'estimation de distance Les techniques d'estimation de distance permettent d'estimer la distance entre deux hôtes dans l'Internet par une mesure de délai. Des propositions tels que *IDMaps* [61], *GNP* (*Global Network Positioning*) [62] ont été développées. Plusieurs systèmes de coordonnées ont été aussi proposés. Ces systèmes ont comme objectif d'associer des coordonnées à chaque machine sur Internet. Ces coordonnées sont construites sur base de mesures faites avec quelques machines voisines et leur intérêt principal est que sur base des coordonnées d'un système distant il est possible de prédire le délai entre le système local et le système distant. Nous pouvons citer des propositions telles que : *King* [63], *Lighthouses* [64], *BBS* (*Bing-Bang Simulation*) [65], *Virtual Landmarks* [66], *ICS* (*Internet Coordinate System*) [67], et *Vivaldi* [68]. Il faut noter que ces différentes techniques d'estimation de distance ne fournissent pas la distance géographique entre deux hôtes, mais plutôt le délai.

2.2.3.4.2 Corrélation entre délai et distance géographique Ballintintijn *et al.* [69] estiment qu'il y a peu de corrélation entre délai et distance géographique. Ce manque de corrélation est attribué à la non linéarité des chemins entre deux noeuds et à la présence de goulets d'étranglements. Toutefois, durant ces toutes dernières années, l'Internet a connu une forte croissance, notamment en terme de bande passante et de couverture. Par exemple, on peut citer le nombre important et la capacité des liens à haut débit, le nombre de points de présence des FAI, ...). Une forte connectivité implique assez souvent plus de linéarité entre les noeuds de bout en bout sur un chemin, mais aussi une forte corrélation entre distance géographique et délai [70]. De plus, des découvertes récentes [21, 22] montrent qu'il existe une forte corrélation entre la population démographique et le nombre de routeurs dans les pays développés. De cet fait, la plupart des hôtes, et par conséquent la majorité des hôtes cibles, se trouvent dans des régions qui possèdent une riche connectivité, et où l'on note une forte corrélation entre délai et distance géographique.

2.2.3.4.3 Inférence de la localisation géographique avec GeoPing La technique GeoPing [52] infère la localisation géographique des hôtes dans l'Internet à partir des mesures de délai. GeoPing mesure le délai vers l' hôte cible et les hôtes références (hôte dont on connaît la position géographique) à partir de plusieurs machines sondes, et combine ces mesures de délai pour inférer la position géographique de cet hôte cible. L'estimation de localisation se base sur l'hypothèse que des hôtes ayant un délai similaire par rapport à d'autres hôtes fixes (des machines sondes par exemple) tendent à être situés dans une même zone géographique. Cette hypothèse est similaire à celle utilisée dans RADAR [34], concernant la relation entre la distance et la force du signal, pour déterminer le positionnement des terminaux mobiles dans un réseau sans fil. Ainsi, la localisation de l'hôte cible est assimilable à la position de l'hôte référence, qui a la mesure de délai la plus similaire par rapport à l'hôte dont on veut déterminer sa localisation. Par conséquent, le nombre et le placement des hôtes références (*landmark*) jouent un rôle important dans la précision de l'estimation de localisation [71].

Soit un ensemble $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ de K hôtes références. Les hôtes références sont des hôtes dont on connaît la position géographique. Soit l'ensemble $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ de N serveurs sondes. La figure (Fig. 2.3) illustre le processus de localisation géographique d'un hôte cible à partir de mesures de délai. Les serveurs sondes font des mesures vers les hôtes références (Fig. 2.3(a)) périodiquement.

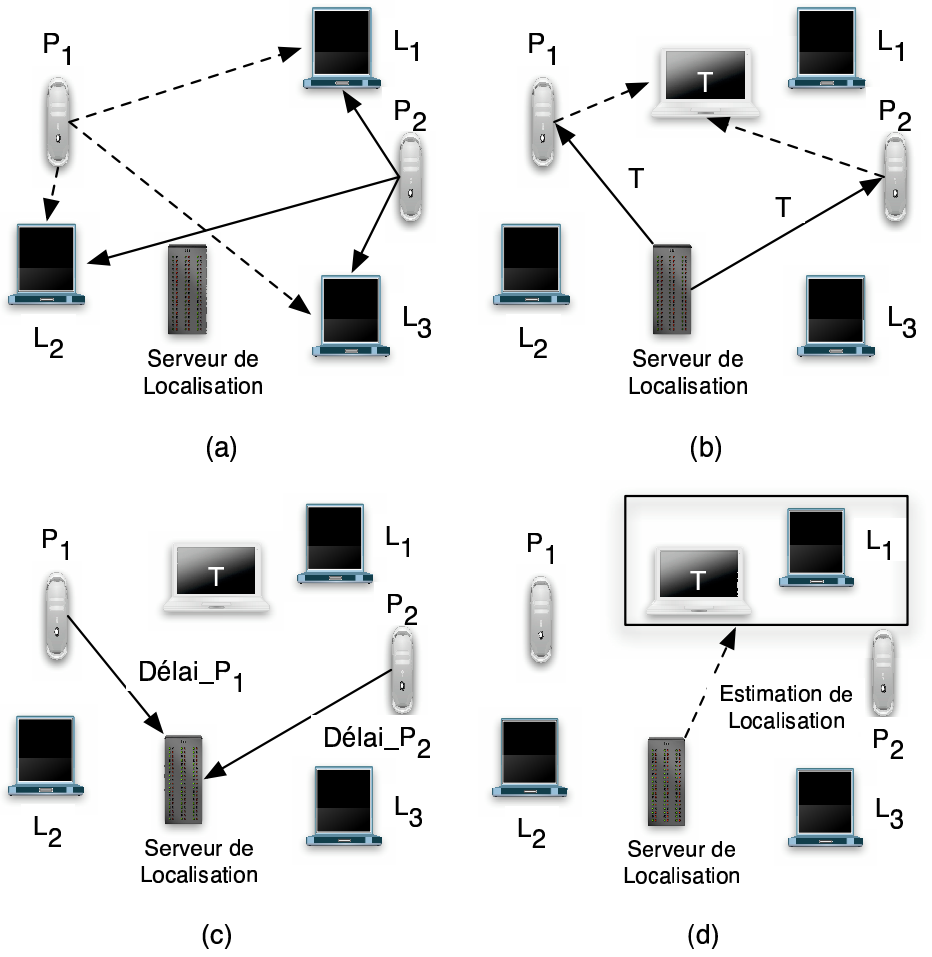


FIG. 2.3 – Inférence de la localisation par GeoPing.

Ainsi, chaque serveur sonde P_x, $1 \leq x \leq N$, retourne un vecteur de délai

$$d_x = (d_{1x}, d_{2x}, \ldots, d_{Kx}), \tag{2.1}$$

où d_{ix} représente le délai minimum entre le serveur sonde P_x et l'hôte référence $L_i \in \mathcal{L}$.

Pour déterminer la localisation d'hôte cible T, un serveur de localisation qui gère l'ensemble des hôtes références $\mathcal{L}$ et des serveurs sondes $\mathcal{P}$ est utilisé. Le serveur de localisation demande aux N serveurs sondes de mesurer le délai vers l'hôte cible T (Fig. 2.3(b)). Chaque serveur sonde P_x, $1 \leq x \leq N$, retourne un nouveau vecteur de délai

$$d'_x = (d_{1x}, d_{2x}, \ldots, d_{Kx}, d_{Tx}), \tag{2.2}$$

lequel est composé du vecteur de délai d_x et du délai vers l'hôte cible T qu'on vient juste de mesurer (Fig. 2.3(c)).

Après avoir reçu les mesures de délai faites par les N serveurs sondes, le serveur de localisation construit la matrice de délai D qui a pour dimension $(K+1) \times N$:

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{K1} & d_{K2} & \dots & d_{KN} \\ d_{T1} & d_{T2} & \dots & d_{TN} \end{pmatrix}$$

Les colonnes de la matrice D représentent les mesures de délai faites par les serveurs sondes vers l'ensemble des hôtes références et la cible T. Pour estimer la localisation géographique de la cible, le serveur de localisation compare les lignes de la matrice D. La technique GeoPing utilise la distance euclidienne pour trouver l'hôte référence qui a la mesure de délai la plus similaire par rapport à l'hôte dont dont on veut déterminer la localisation. Par exemple, l'hôte référence L_y ayant la plus petite distance euclidienne

$$e_{L_yT} = \sqrt{(d_{y1} - d_{T1})^2 + (d_{y2} - d_{T2})^2 + \ldots + (d_{yN} - d_{TN})^2}, \tag{2.3}$$

par rapport à l'hôte T où $y = 1, \ldots, K$, donne sa position comme estimation de localisation de la cible (Fig. 2.3(d)).

Le nombre d'endroits possibles, où on peut localiser l'hôte cible, est ainsi limité au nombre d'hôtes références composant notre système. Nous obtenons un espace discret de réponses. Ainsi, la précision de l'estimation de localisation dépend du nombre et du placement des hôtes références [71]. Cependant, augmenter le nombres d'hôtes références conduit à accroître le nombre de mesures pour inférer la localisation des hôtes cibles. Le nombre de mesures $\mathcal{M}_{(n,\tau)}$ nécessaire pour estimer la localisation géographique de n cibles dans un intervalle de temps τ peut être modélisé comme suit :

$$\mathcal{M}_{(n,\tau)} = 2N\left(\left\lceil \frac{\tau}{\Delta} \right\rceil K + t\right). \tag{2.4}$$

où Δ représente la périodicité des mesures faites par les serveurs sondes vers l'ensemble des hôtes références $\mathfrak{L}$. Cette périodicité des mesures, entre les serveurs sondes et les hôtes références, permet d'obtenir une matrice de délai représentative dans le temps. Par exemple, des mesures faites en jours ouvrables ou bien durant la nuit d'un week end peuvent être différentes.

Le facteur multiplicatif 2, dans l'équation 2.4, est dû au fait que chaque mesure de délai, faite avec l'outil ping, se compose de deux messages. Ainsi, nous avons

un message `ECHO_REQUEST` et un message `ECHO_RESPONSE`. Il faut aussi noter que chaque mesure de délai est constitué d'une ou de plusieurs échantillons et seul l'échantillon de mesures ayant le plus petit délai est considéré. Dans le cas où on veut évaluer le RTT minimum, entre chaque serveur sonde et les hôtes références, avec un ping composé de p paquets, le trafic injecté dans le réseau est donné par $p \times \mathcal{M}$. Ainsi, augmenter le nombre d'hôtes références revient à surcharger le réseau.

Dans [71], les auteurs proposent d'utiliser d'autres modèles de similarité [72] autres que la distance euclidienne, comme GeoPing le fait, pour trouver l'hôte référence qui présente la mesure de délai la plus similaire par rapport à l'hôte cible. Les résultats obtenus dans [71] montrent que des modèles de similarité tels que "*distance-based*" ou " *city-block distance*" ont une meilleure performance que la distance euclidienne. Nous obtenons ainsi une meilleure estimation de localisation des hôtes Internet avec ces nouvelles modèles de similarité.

2.2.3.4.4 Placement démographique des hôtes références et des serveurs sondes Pour améliorer la précision de la technique GeoPing, [73, 74] proposent un placement démographique, des hôtes références et des serveurs sondes, qui tient compte de la répartition géographique des utilisateurs. L'auteur considère toutes les agglomérations qui ont plus d'un million d'habitants [75]. Comme les infrastructures qui composent l'Internet sont réparties de manière disproportionnées à travers les différentes régions du monde, [74] pondère les populations des différentes agglomérations par leur nombre d'internautes. Le rapport entre le nombre total d'internautes dans le pays [76] sur sa population totale donne cette pondération. En appliquant cette pondération, les principales agglomérations qui doivent recevoir les hôtes références, sont ainsi définies. Le placement des hôtes références et des serveurs sondes se fait suivant les approches proposées dans [77, 78, 79] pour le placement des infrastructures urbaines.

Les résultats obtenus montrent que le placement démographique des hôtes références permet d'utiliser un nombre restrictif d'hôtes références capable de représenter une portion importante des utilisateurs avec une limite de distance fixe. L'utilisation d'un nombre limité d'hôtes références permet de réduire aussi le volume de mesures injectées dans le réseau. Le fait aussi de placer de manière démographique les serveurs sondes permet d'éviter une redondance au niveau des mesures.

2.2.3.4.5 Placement hiérarchique des hôtes références Un placement hiérarchique à deux niveaux des hôtes références est aussi proposé dans [80] afin de diminuer le nombre de mesures générées par les serveurs sondes vers les hôtes références pour localiser un hôte cible. Ce placement hiérarchique permet de localiser dans un premier temps l'hôte cible par les hôtes références situés au niveau

TAB. 2.1 – Notation pour une structure hiérarchique à q niveaux.

$\mathfrak{L}_s^q$	hôtes références du sous-ensemble s de niveau q
h_{is}^q	nombres d'utilisateurs de l'agglomération i couverts par le sous-ensemble s de niveau q
U_s^q	nombre d'utilisateurs couverts par le sous-ensemble s de niveau q ; $U_s^q = \sum_j h_{sj}^q$
K_s^q	nombre d'hôtes références du sous-ensemble s de niveau q
a_{ij}^q	$\begin{cases} 1 & \text{si l'agglomération } i \text{ couvre l'agglomération } j \text{ dans le niveau } q \\ 0 & \text{sinon.} \end{cases}$
X_i^q	$\begin{cases} 1 & \text{si un hôte référence est placé dans l'agglomération } i \text{ du niveau } q, \\ 0 & \text{sinon.} \end{cases}$
Z_i^q	$\begin{cases} 1 & \text{si l'agglomération } i \text{ de niveau } q \text{ est couverte,} \\ 0 & \text{sinon.} \end{cases}$
c_s	$\begin{cases} 1 & \text{si } \mathfrak{L}_s^q \subset \mathfrak{L}_j^{q-1}\ \forall j, \\ 0 & \text{sinon.} \end{cases}$

supérieur, puis de raffiner, si nécessaire, l'estimation de localisation par les hôtes références situés aux niveaux inférieurs.

Dans le premier niveau de de la structure hiérarchique, nous plaçons un nombre réduit d'hôtes références de manière très espacée les uns des autres. Chaque hôte référence couvre une très grande zone géographique, et l'estimation de localisation se fait de manière assez générale. L'ensemble des hôtes références placés au deuxième niveau est obtenu en diminuant la distance de couverture de ceux placés au premier niveau. Ainsi, le deuxième niveau fournit plus de précision pour l'estimation de la localisation.

Nous adoptons la notation définie dans le tableau 2.1 pour montrer la distribution des hôtes références dans une structure hiérarchique à q niveaux.

Cette structure hiérarchique menant aux sous-ensembles $\mathfrak{L}_s^q$ de l'ensemble des hôtes références $\mathfrak{L}$ est obtenue par

$$\mathfrak{L} \supseteq \bigcup_{s,q} \mathfrak{L}_s^q. \tag{2.5}$$

Comme illustré sur la figure (Fig. 2.4), le niveau supérieur de notre structure hiérarchique ($q = 1$) contient un ensemble $\mathfrak{L}_1^1$ composé de K_1^1 hôtes références. Par contre le niveau inférieur ($q = 2$) contient les ensembles $\mathfrak{L}_s^2$ avec chacun respectivement K_s^2 hôtes références, pour tout s.

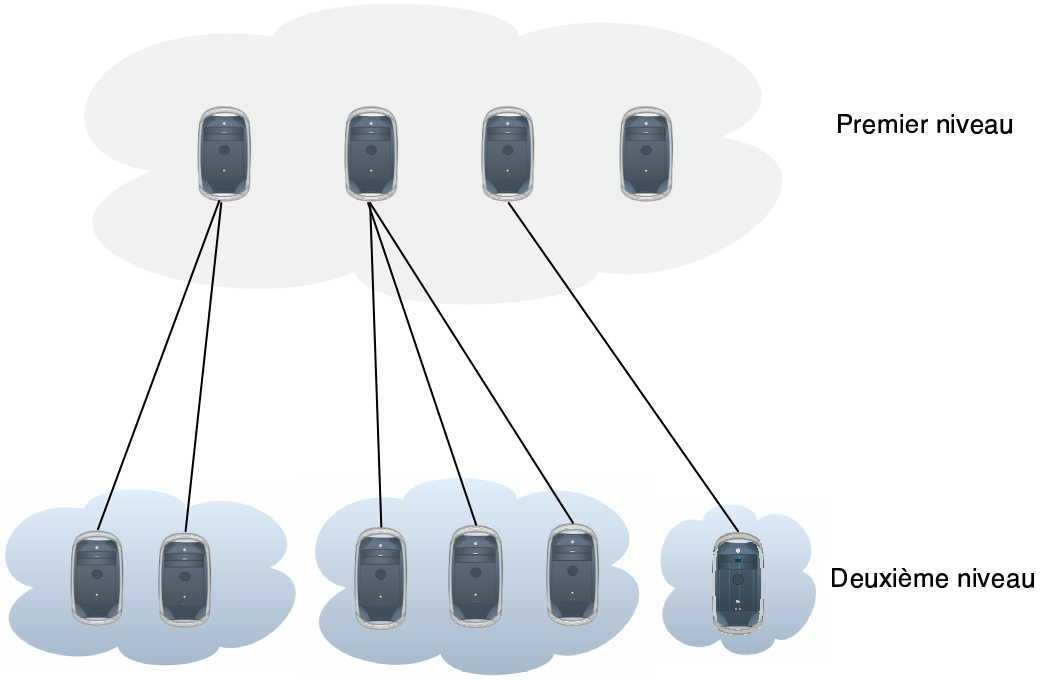


FIG. 2.4 – Exemple d'une structure hiérarchique à deux niveaux.

Les hôtes références sont placés en fonction de la distribution géographique des utilisateurs (hôtes) à travers le monde suivant les critères proposés dans [73] pour un placement à un seul niveau des hôtes références. L'idée de ce placement démographique par politique est de placer les hôtes références dans les localités où ils sont censés couvrir le plus d'utilisateurs possible. En utilisant les notations du tableau 2.1, cette approche est formulée par la fonction

$$\max \sum_{q} \sum_{s} \sum_{i} h_{is}^{q} Z_{i}^{q}. \tag{2.6}$$

Cette équation maximise le nombre d'utilisateurs couverts par un hôte référence situé au niveau q. Cette fonction est assujettie aux contraintes suivantes :

$$Z_{i}^{q} \leq \sum_{j} a_{ij}^{q} X_{j} \qquad \forall i, q \tag{2.7}$$

$$\sum_{j} X_{j}^{q} = K_{s}^{q} \qquad \forall s, q \tag{2.8}$$

$$Z_{i}^{q} = 0, 1 \qquad \forall i, q \tag{2.9}$$

$$X_{i}^{q} = 0, 1 \qquad \forall i, q \tag{2.10}$$

Le nombre de mesures générées dans le réseau par cette structure hiérarchique pour localiser un hôte cible, dépend de la localisation géographique de cet hôte et du degré de précision que nécessite l'estimation de localisation. Si certaines agglomérations possèdent plus d'utilisateurs que d'autres, les hôtes références s'y trouvant vont être plus solliciter pour fournir une localisation, car la demande à localiser un hôte y est beaucoup plus fréquente. Ainsi, il y a certaines agglomérations où l'on note beaucoup plus de mesures que dans d'autres. Supposons que chaque hôte ait la même probabilité d'être localiser, cette probabilité p_s^q donnée par

$$p_s^q = \frac{\sum_{j=1}^{K_s^q} h_{js}^q}{\sum_{j=1}^{K_s^{q-1}} h_{js}^{q-1}}, \quad \forall s, \tag{2.11}$$

exprime la probabilité de faire des mesures vers les hôtes références du sous-ensemble s localisé au niveau q. Dans notre structure hiérarchique où q= 2, nous avons la probabilité p_s^2 pour chaque hôte référence de l'ensemble K_1^1 composant le niveau supérieur. Par exemple, la probabilité p_1^2 définit la probabilité que l'on a pour utiliser les hôtes références du premier sous-ensemble ($i = 1$) localisé au niveau inférieur ($q = 2$) pour améliorer l'estimation de localisation. Donc pour l'approche à deux niveaux proposée, où nous avons qu'un ensemble d'hôtes références au premier niveau, la probabilité p_s^2 s'écrit comme suit :

$$p_s^2 = \frac{\sum_{j=1}^{K_s^2} h_{js}^2}{\sum_{j=1}^{K_1^1} h_{js}^1} = \frac{U_s^2}{U_1^1}, \quad \forall s. \tag{2.12}$$

Connaissant la probabilité qui existe pour faire des mesures vers les hôtes références localisés dans chacun des sous-ensembles du second niveau, nous évaluons la charge du réseau de cette structure hiérarchique à deux niveaux. Le nombre minimum de mesures générées dans le réseau pour localiser un hôte t est

$$\mathcal{M}_{\text{2niveaux}}^{\min}(t) = 2N \left[(K_1^1 + t) + \sum_{s=1}^{K_1^1} (K_s^2 - c_s) p_s^2 \right], \quad \forall i. \tag{2.13}$$

Il est important de noter que les sous-ensembles regroupant les hôtes références ne sont pas forcement disjoints. De même un hôte référence appartenant à l'ensemble $\mathcal{L}_1^1$ et couvrant ainsi une large zone géographique dans le niveau supérieur de notre structure hiérarchique à deux niveaux peut appartenir à un autre sous-ensemble du niveau inférieur. Si un hôte référence est localisé simultanément dans les deux niveaux, il n'est pas nécessaire de faire deux mesures vers ce même hôte référence si on devait toutefois améliorer l'estimation de localisation. Le pa-

Algorithm 1 Approche du MCLM [78]

1 : $\mathcal{L} \leftarrow \emptyset$; $\mathcal{A}' \leftarrow \mathcal{A}$
2 : **tant que** ($|\mathcal{L}| < K$) **faire**
3 : Trouver $A \in \mathcal{A}'$ qui couvre le plus d'utilisateurs non encore couverts
4 : attribuer $\mathcal{C} \subseteq \mathcal{A}'$ comme l'ensemble des agglomérations couvertes par A
5 : $\mathcal{L} \leftarrow \mathcal{L} \cup A$; $\mathcal{A}' \leftarrow \mathcal{A}' - \mathcal{C}$
6 : **fin tant que**
7 : $\mathcal{L}$ est l'ensemble qui contient les K hôtes références

ramètre c_s permet d'éviter qu'une deuxième mesure soit faite vers le même hôte référence si au premier niveau cette mesure fut déjà réalisée.

2.2.3.4.5.1 Mise en place des deux niveaux hiérarchiques Dans le premier niveau de notre structure hiérarchique le nombre d'hôtes références utilisés est limité. Nous fixons la distance de couverture assignée à chaque hôte référence et maximisons le nombre d'hôtes couverts par chacun d'eux. Cette approche est appelée MCLM (*Maximum Covering Location Model*) [78]. Nous adoptons l'algorithme 1 pour résoudre le problème du MCLM. Cet algorithme possède une complexité d'ordre $O(|\mathcal{A}|^2 K)$. Il permet de placer des hôtes références qui couvrent le plus d'utilisateurs non encore couverts jusqu'à ce que K hôtes références soient placés. Ainsi, nous déterminons l'ensemble $\mathcal{L}_1^1$ des hôtes références du premier niveau qui contient K_1^1 hôtes références.

Le deuxième niveau de notre structure hiérarchique est obtenu en réduisant les distances de couverture des hôtes références localisés au niveau supérieur. En effet les hôtes références du premier niveau couvrent une large zone géographique regroupant un très grand nombre d'utilisateurs. Ainsi pour chaque zone du premier niveau couverte par un hôte référence, nous plaçons des hôtes références supplémentaires avec un rayon de couverture beaucoup plus petit. Ces hôtes références couvrent toutes les agglomérations qui sont situées dans cette zone géographique. Nous déterminons ainsi les ensembles $\mathcal{L}_s^q$ d'hôtes références qui composent le deuxième niveau de notre structure hiérarchique.

Les résultats obtenus dans [80] montrent que la structure hiérarchique à deux niveaux réduit très largement le nombre de mesures effectuées pour estimer la position d'un hôte cible comparé aux mesures que génèrent la technique Geo-Ping [52] (structure plate). Par contre, le pourcentage d'utilisateurs couverts dans la structure hiérarchique diminue par rapport au pourcentage d'utilisateurs couverts dans la structure plate pour le même nombre d'hôtes références. Cependant les résultats démontrent que la diminution de la charge du réseau compense la diminution du nombre d'utilisateurs couverts. La structure hiérarchique diminue ainsi l'impact des mesures dans le réseau mais reste toutefois confrontée à la

précision des mesures. En effet, le nombre d'endroits possibles, où on peut localiser un hôte cible, est limité au nombre d'hôtes références (espace discrets de réponses).

2.2.3.4.6 Octant Tout récemment, une technique de géolocalisation nommée *Octant* [81], a été développée. Cette technique se base sur l'utilisation de la multilatération dans l'Internet [26]. Ainsi, Octant infère une zone géographique β_i, délimitée par un ensemble de points se trouvant sur la surface du globe, dans laquelle l'hôte cible i pourrait être localisé. L'estimation de la zone géographique β_i est faite en utilisant les contraintes $\gamma_0 \dots \gamma_n$. En effet, une contrainte γ est une région du globe, dans laquelle l'hôte cible est supposé être, et à laquelle on associe une pondération pour évaluer la confiance en cette région.

Les contraintes sont obtenues à partir des mesures de délai faites par les hôtes références ("*landmarks*"), hôtes dont on connaît la position géographique, et qui sont choisis au hasard à partir de l'ensemble des clients. A chaque hôte référence L_j, on associe une région β_{L_j} dont la superficie capture l'incertitude que l'on a par rapport à l'estimation de la position de l'hôte référence. Les auteurs de Octant [81] appellent "*primary landmark*" tout nœud dont la position est connue à partir des mesures faites par un GPS ou à partir du mappage d'une zone géographique en coordonnées géographiques. En d'autres mots, c'est un hôte référence dont la position est connue avec précision. Quant à "*secondary landmark*", c'est tout nœud dont la position géographique a été estimée par l'intermédiaire de l'outil Octant. Dans de tels cas, la zone géographique β_{L_j} est obtenue en exécutant Octant avec le secondary landmark L_j (hôte référence) choisi comme cible.

L'outil Octant permet aux hôtes références (landmarks) d'inférer la position géographique des hôtes cibles en introduisant deux types de contraintes : une frontière extérieure (contrainte positive) qui détermine que la distance maximale entre la cible τ et l'hôte référence L_1, et une frontière intérieure (contrainte négative) qui détermine la distance minimale entre la cible τ et l'hôte référence L_1. Si la position géographique du primary landmark est connue, nous sommes dans le cas où nous avons une contrainte positive ("*positive constraint*") avec une distance d qui définit un disque de rayon d et ayant pour centre la position de l'hôte référence (primary landmark). La position de l'hôte cible est inférée à l'intérieur de ce disque. Par contre, une contrainte négative ("*negative constraint*") avec une distance d', définie la zone formée par l'ensemble des points qui sont à l'extérieur du disque de rayon d'. Quand l'hôte référence considéré est un primary landmark (*i.e* hôte référence dont la position géographique est connue avec exactitude), de telles contraintes définissent un anneau.

Pour un secondary landmark k (*i.e* hôte référence dont la position géographique est déterminée avec Octant), dont la zone géographique estimée est β_k, une

contrainte positive avec un rayon d définie une région formée par l'union de tous les cercles de centre l'ensemble des points intérieurs à β_k et de rayon d. Cette contrainte est formalisée par

$$\gamma = \bigcup_{(x.y)\in\beta_k} c(x, y, d), \tag{2.14}$$

où c(x,y,d) représente le disque de rayon d centré en (x,y).

Par contre, une contrainte négative élimine la possibilité de localiser l'hôte cible au niveau des points qui sont à une distance d de l'hôte référence, et indépendamment de sa position à l'intérieur de la zone β_k. Cette contrainte négative est formalisée par

$$\gamma = \bigcap_{(x.y)\in\beta_k} c(x, y, d), \tag{2.15}$$

où c(x,y,d) représente le disque de rayon d centré en (x,y). L'utilisation des courbes de *Bézier* par Octant permet de représenter efficacement ces régions en transformant uniquement les points qui se trouvent aux extrémités des segments de Bézier.

Soit Ω l'ensemble des contraintes positives et Φ l'ensemble des contraintes négatives associés à la position de l'hôte cible i. L'hôte cible i se trouve dans la région définie par

$$\beta_i = \bigcap_{X_i\in\Omega} X_i \bigcup_{X_i\in\Phi} X_i. \tag{2.16}$$

La figure (Fig. 2.5) illustre comment Octant infère la zone géographique dans laquelle l'hôte cible se trouve. En effet, Octant calcule la zone β en combinant les contraintes positives et négatives obtenues à partir des mesures délai faites par les hôtes références vers l'hôte cible. La zone géographique estimée est formée par des régions non convexes, potentiellement disjointes séparées par des pondérations (voir Fig. 2.5). La zone géographique choisie comme l'estimation de la cible, est l'union des contraintes ayant la plus grande pondération. Pour donner une estimation de localisation dans cette zone géographique, Octant utilise l'algorithme de *Monte-Carlo* pour choisir un point qui représente la meilleur estimation de localisation de la cible dans cette zone.

Dans [81] il est aussi proposé d'utiliser les routeurs intermédiaires sur le chemin entre un hôte référence et un hôte cible comme hôtes référence additionnels. Octant utilise la base de données de *udns* [82] pour déterminer la position géographique de plusieurs routeurs intermédiaires. Pour les routeurs qui n'ont pas pu être localisés par la base de données udns, Octant propose d'utiliser leur estimation de localisation.

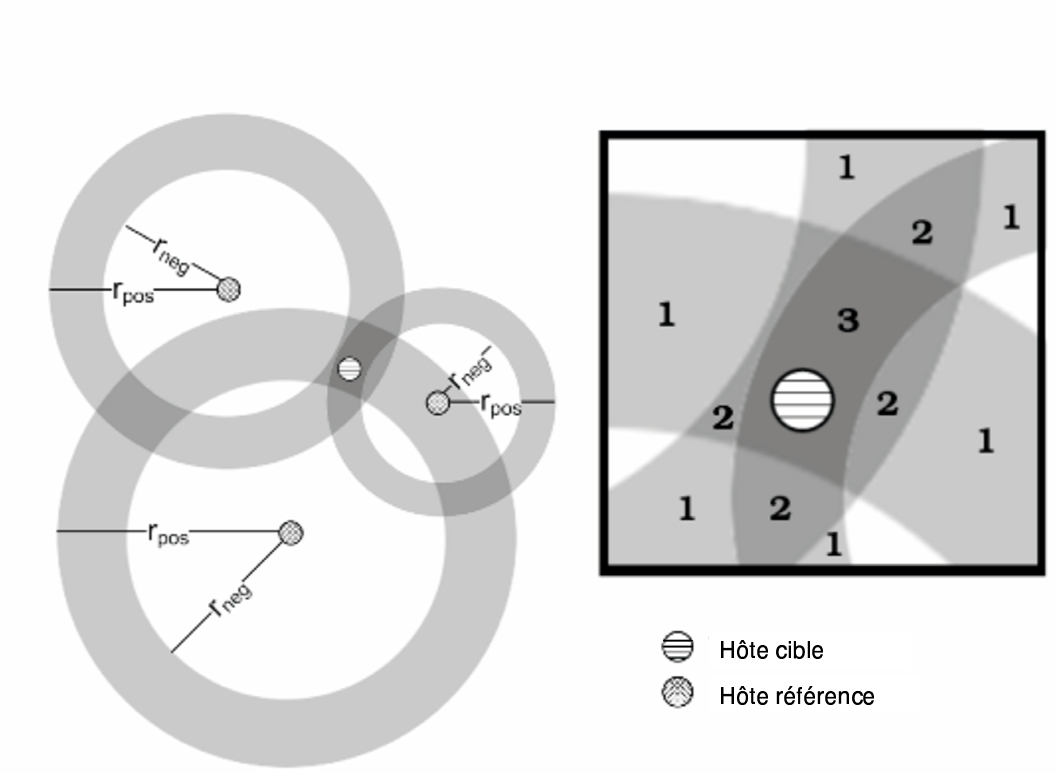


FIG. 2.5 – Estimation de la localisation d'un hôte cible avec Octant.

Toutefois, l'outil Octant n'est pour le moment disponible que pour la localisation des hôtes situés aux États Unis.

2.2.4 Conclusion

La localisation géographique peut être appliquée dans des domaines tels que : le GSM, les réseaux ad hoc, et dans l'Internet. Toutefois, nous nous focalisons sur la localisation géographique des hôtes dans l'Internet à partir uniquement de leur adresse IP. La localisation géographique d'un hôte dans l'Internet peut être obtenue grâce à l'utilisation de base de données Whois mais aussi par le biais du nom DNS de l'hôte cible ou des routeurs qui lui sont proches. Pour les techniques basées sur les noms DNS, nous avons vu que tous les routeurs ne possédaient pas un nommage conventionnel, d'où une difficulté pour extraire l'information de localisation. De plus, une mauvaise estimation peut être obtenue si le client se connecte via un proxy et/ou un firewall, ou bien si le routeur qui donne sa position comme estimation de la cible est assez éloigné.

Basée sur le clustering, la technique GeoCluster utilise les informations des

tables de routage BGP et une base de données contenant des adresses IP et leur localisation. Bien que GeoCluster ne génère aucune mesure dans le réseau, son efficacité dépend de la véracité et de la représentativité des informations, fournies par les utilisateurs, se trouvant dans la base de données IP-Localisation.

Quant aux techniques basées sur les mesures de délai, elles s'appuient sur une corrélation entre distance géographique et délai. GeoPing choisit la position de l'hôte référence (landmark), hôte dont on connaît la position géographique, qui a la mesure de délai la plus similaire par rapport à l'hôte cible comme son estimation de localisation. Ainsi le nombre d'endroits possibles où on peut localiser un hôte cible, est limité au nombre d'hôtes références. Nous obtenons un espace discret de réponses. Par conséquent, le nombre d'hôtes références joue un rôle important dans la précision de l'estimation. Vu le nombre de messages à envoyer pour localiser une cible, on ne peut augmenter le nombre d'hôtes références de manière incontrôlée. En outre, l'hôte référence qui donne sa position comme estimation de localisation de la cible peut être assez éloigné, d'où une mauvaise estimation.

Pour remédier à ces verrous, nous proposons la technique *CBG*, basée sur la *multilatération*, pour inférer la position géographique d'un hôte cible. La multilatération permet d'obtenir un espace continu d'endroits possibles où on peut localiser un hôte cible. Elle permet également d'associer une zone de confiance à chaque estimation de localisation.

Chapitre 3

Localisation géographique basée sur la Multilatération

Dans ce chapitre, nous proposons et décrivons la technique *CBG* (*Constraint-Based Geolocation*) pour remédier aux limitations des techniques basées sur les mesures de délai. Nous montrons comment CBG transforme, avec des contraintes, les mesures de délai en distances géographiques estimées (*geographic distance constraints*) avant d'appliquer la *multilatération* pour inférer la position géographique des hôtes cibles. Nous montrons également que CBG associe à chaque estimation de localisation une zone de confiance.

3.1 Introduction à CBG

Les techniques précédentes de localisation géographique, basées sur les mesures de délai [71, 52], utilisent la position de l'hôte référence (*landmark*) le plus proche, en terme de délai, comme possible localisation de l'hôte cible. On appelle hôte référence tout hôte dont la position géographique est connue. Avec cette approche, l'ensemble des endroits où on peut localiser un hôte cible est limité par le nombre d'hôtes références. Nous obtenons ainsi un ensemble discret de réponses formé par l'ensemble des endroits où sont localisés nos hôtes références. Pour surmonter cette limitation, nous proposons l'approche CBG [27, 26] qui préconise d'utiliser la multilatération. A ma connaissance, CBG est la première technique de géolocalisation à utiliser la multilatération pour inférer la position des hôtes dans l'Internet.

En effet, la multilatération permet d'estimer une position en utilisant un nombre suffisant de distances à partir de quelques points immobiles. Dès lors, elle fournit un ensemble continu d'endroits où on peut localiser la cible au lieu

d'un espace discret de réponses. Nous utilisons un ensemble d'hôtes références pour estimer la localisation des cibles. En appliquant la multilatération avec la "*distance géographique estimée*" entre la cible et chaque hôte référence, CBG fournit une estimation de localisation de l'hôte cible, comme le fait le système de positionnement par satellites (GPS) [16] en considérant la distance entre chaque satellite et la cible. Toutefois pour pouvoir appliquer la multilatération dans l'Internet, il faut que les distances géographiques utilisées soient obtenues à partir des mesures de délai. Ceci est un véritable défi car la distance géographique n'est pas fortement corrélée avec le délai [69]. Cela est dû aux congestions qui existent dans les réseaux, ajoutant un délai supplémentaire dans les mesures, la violation de la règle de l'inégalité triangulaire [23], et à la non linéarité des chemins entre les hôtes [24].

L'élément clé de CBG est sa capacité à transformer les mesures de délai en distances géographiques surestimées. Sachant que les informations se propagent dans une fibre optique à une vitesse équivalente à 2/3 la vitesse de la lumière dans le vide [83], à partir de toute mesure de délai faite entre deux points, nous pouvons calculer une distance qui est une *borne supérieure* de la distance mesurée entre ces deux points. Cette distance, qui représente une borne supérieure, est égale au délai mesuré divisé par la vitesse de propagation de la lumière dans la fibre. En se basant sur ce raisonnement, pour toute transmission de données entre deux points, il existe un délai minimum théorique dépendant de la distance maximale obtenue précédemment. Ainsi, au délai effectif mesuré s'ajoute un délai supplémentaire dû aux distorsions (congestions, violation de l'inégalité triangulaire, non linéarité des chemins entre les hôtes,...).

Par conséquent, pour obtenir une estimation de localisation précise, il faudra essayer d'évaluer et d'enlever si possible le délai supplémentaire qui s'ajoute aux mesures de délai. En effet, si la technique CBG utilisée directement ces mesures de délai pour inférer l'estimation de localisation de l'hôte cible, elle ne serait pas assez performante en terme de précision. CBG tente d'enlever ces distorsions grâce à un auto-calibrage des mesures de délai obtenues entre les hôtes références (voir section 3.3). Après l'auto-calibration, CBG peut ainsi transformer, avec une certaine précision, les mesures de délai entre les hôtes références et la cible en distances géographiques surestimées. L'auto-calibration essaye de contrecarrer les causes du délai supplémentaire qui s'ajoute au délai mesuré. Chaque distance géographique obtenue, à partir de la transformation du délai en distance géographique, est formée par la distance géographique réelle entre l'hôte référence et l'hôte cible, plus la distance induite par le délai supplémentaire qui s'ajoute aux mesures de délai. Cest la raison pour laquelle, nous appelons la distance géographique obtenue distance géographique surestimée. CBG applique la multilatération, en utilisant ces distances géographiques surestimées, pour inférer la zone géographique dans

laquelle se trouve l'hôte cible. Ensuite, CBG considère le centre de cette zone comme une estimation de localisation de l'hôte cible.

Contrairement aux techniques précédentes de localisation géographique, CBG est capable de fournir une zone de confiance pour chaque hôte localisé. Cela permet aux applications, qui utilisent CBG, d'évaluer la fiabilité de l'estimation par rapport à leurs exigences. Ainsi, chaque application peut définir le niveau de précision qu'elle désire. En outre, un serveur de localisation qui implémente CBG peut recevoir une requête de demande de localisation de la part d'un hôte cible, mais aussi de la part d'un serveur Web qui désire localiser ses propres clients.

3.2 La Multilatération : idée générale

La position physique d'un point quelconque peut être estimée en utilisant un nombre suffisant de distances ou d'angles par rapport à des points immobiles, dont on connaît la localisation. Ainsi, la multilatération est la détermination de la position par mesures des distances à plusieurs points de référence. Elle peut être à géométrie sphérique, basée sur l'intersection de sphères correspondant à des mesures de distances, et dont les systèmes GNSS (*Global Navigation Satellite System*), par exemple GPS, GLONASS, Galileo, sont les exemples classiques. Elle peut être aussi à géométrie hyperbolique, mettant en œuvre l'intersection d'hyperboloïdes correspondant à des mesures de différences de distances. Par contre, si des angles sont choisis, l'approche est appelée multiangulation. Assez souvent, le terme triangulation est utilisé pour l'estimation d'un point à partir de mesures de distances ou d'angles. Toutefois la triangulation se définit comme une méthode trigonométrique permettant de déterminer la position d'un point fixe, en utilisant les angles mesurés entre ce point et trois autres points fixes considérés comme points références. Le terme multilatération, étant celui le plus approprié, est utilisé dans le reste de cette thèse au détriment du terme triangulation, malgré sa popularité. Nous appliquons également la multilatération à géométrie sphérique.

La multilatération exige une précision accrue des mesures de distance entre l'hôte cible et les hôtes références. Par exemple, le système de positionnement par satellites (GPS) [16] utilise la multilatération pour localiser un récepteur GPS. On mesure la distance entre le récepteur et trois satellites dont les positions sont connues. Pour ce faire, le récepteur mesure le temps mis par le signal du satellite pour lui parvenir et le convertit en distance. La synchronisation de l'horloge du récepteur avec le GPS et la précision des mesures sont les éléments principaux qui font que le système GPS est précis. Chaque satellite est muni d'une horloge atomique à son bord. Dans le cas du GPS, la multilatération étant effectuée avec des distances presque exactes (obtenues à partir du temps mis par le signal), l'estimation de localisation est alors faite avec précision.

Par contre, la transformation des mesures de délai en distances géographiques avec précision est un véritable défi dans l'Internet. C'est la raison pour laquelle la multilatération est restée inutilisée jusqu'à présent pour inférer la localisation géographique des hôtes dans l'Internet. Par la suite, nous expliquons la conception de la technique CBG, et comment elle applique la multilatération en considérant des distances géographiques surestimées.

Soit un ensemble $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ de K hôtes références (hôte dont on connaît la position géographique). Connaissant le délai entre un hôte cible et l'ensemble des hôtes références, notre principal objectif est d'estimer les distances géographiques correspondantes. En effet, le délai de bout en bout peut être divisé en deux composants : le délai déterministe (ou fixe) et le délai stochastique [84]. Le délai déterministe est composé du temps de traitement au niveau de chaque routeur, du temps de transmission et du délai de propagation. Le délai stochastique est composé du temps d'attente dans les files des routeurs et du temps de traitement au niveau de chaque routeur étant supérieur au temps minimum de traitement. Au delà du délai stochastique, la transformation du délai en distance géographique peut souffrir d'autres formes de distorsions. L'effet de ces distorsions sur la relation entre délai et distance géographique est évalué dans le chapitre 4. Au délai effectif mesuré s'ajoute donc un délai supplémentaire induit par ces distorsions. Par conséquent, l'estimation de distance fournie par CBG est composée de la distance géographique réelle à laquelle s'ajoute une distance induite par ces distorsions.

L'idée fondamentale sur laquelle se base CBG, est qu'au délai mesuré s'ajoute un temps supplémentaire par rapport au temps de propagation de la lumière sur la fibre sur un chemin entre deux nœuds de bout en bout. En se basant sur ce fait, nous avons développé une méthode qui infère la *distance géographique surestimée* à partir du délai mesuré. Il faut noté qu'un délai supplémentaire, dû aux sources de distorsions, s'ajoute à ce délai mesuré. Dans la section 3.3, nous montrons comment CBG infère la distance géographique surestimée entre les hôtes références et l'hôte cible à partir de leur mesures de délai respectives. Il est aussi montré que la présence de ce délai additionnel, induit généralement une surestimation de la distance géographique estimée par rapport à la distance géographique réelle.

La figure (Fig. 3.1) montre le principe de la multilatération en considérant un ensemble $\mathcal{L} = \{L_1, L_2, L_3\}$ d'hôtes références. Chaque hôte référence L_i tente d'inférer sa distance géographique surestimée par rapport à l'hôte cible τ dont la position géographique est inconnue. Toutefois , la distance géographique inférée est donnée par $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$, où $g_{i\tau}$ représente la distance géographique réelle et $\gamma_{i\tau}$ une distance géographique additionnelle. Cette distance géographique additionnelle $\gamma_{i\tau}$ provient du délai supplémentaire, dû aux distorsions et, imbriqué dans le délai de bout en bout. La présence de cette distance géographique ad-

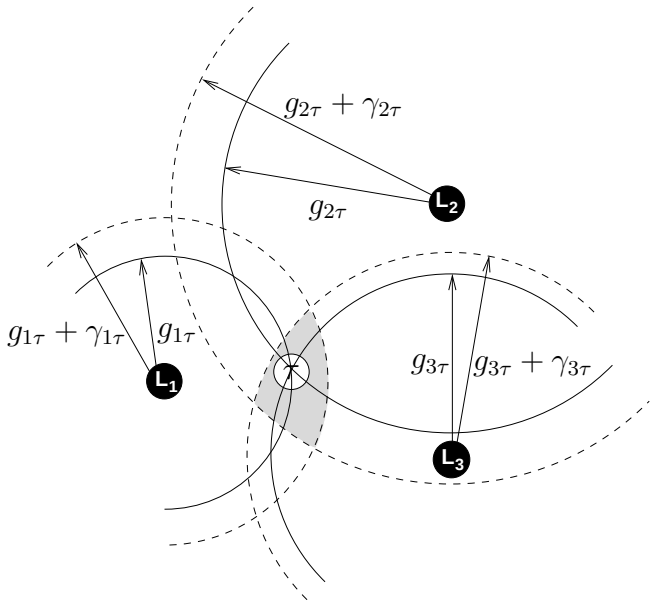


FIG. 3.1 – Multilatération utilisant des distances géographiques surestimées.

ditionnelle permet d'obtenir la zone grise (*voir* Fig. 3.1). L'hôte cible se trouve quelque part dans cette zone. Cette zone grise correspond à la zone d'intersection des cercles, ayant pour centre la position géographique de chaque hôte référence et pour rayon la distance géographique surestimée entre cet hôte référence et l'hôte cible.

3.3 Transformations des mesures de délai en distances géographiques surestimées

Avant d'introduire comment CBG convertit les mesures de délai en distances géographiques surestimées, regardons d'abord la relation pouvant existée entre délai et distance géographique. La figure (Fig. 3.2) illustre un exemple choisi parmi les résultats décrits dans le chapitre 4. L'axe des abscisses représente la distance géographique réelle et l'axe des ordonnées le délai mesuré entre un hôte référence L_i et les autres hôtes références restants dans notre ensemble d'hôtes références considéré. Les droites "*bestline*" et "*baseline*" illustrées sur la figure (Fig. 3.2) seront expliquées le long de cette section.

Des travaux récents [71, 85, 52] utilisent la méthode des moindres carrés pour trouver la relation existant entre distance géographique et délai. Cette méthode permet à partir d'un nuage de points de trouver l'équation de la droite qui ajuste

au mieux l'ensemble des points du nuage. Cependant, vu comment les points sont dispersés au niveau de la figure (Fig. 3.2), nous pensons que cette droite ne traduit pas au mieux la relation entre distance géographique et délai. Ainsi, la droite qui capture le mieux une relation pouvant exister entre distance géographique et délai est, la droite la plus proche, mais en dessous de tous les points [86].

En se basant sur ces considérations, nous proposons une nouvelle approche qui établit une relation dynamique entre délai et distance géographique. Supposons qu'il existe un chemin linéaire entre l'hôte référence L_i et tous les autres hôtes références restants, et que les données sont contraintes à aucun facteur à part le délai de propagation sur le support. Dans ce cas idéal, nous devrions avoir une droite de la forme

$$y = mx + b, \tag{3.1}$$

où $b = 0$ puisqu'il n'y a pas de délai additionnel et m n'est rien d'autre que la vitesse de transmission des données dans le support physique.

Sachant que les informations se propagent dans la fibre optique à une vitesse équivalente à 2/3 la vitesse de la lumière dans le vide [83], alors nous obtenons la règle de conversion suivante : 1 ms RTT correspond à 100 km. Cette relation d'équivalence permet d'obtenir le temps minimum de propagation de l'information entre des sites dont leur localisation géographique est connue. Ce temps minimum est représenté par la "baseline", que nous pouvons qualifier de *droite théorique*, montrée sur la figure (Fig. 3.2). Si nous avions ce cas idéal, cette relation entre délai et distance géographique aurait permis de convertir de manière exacte les mesures de délai en distance géographique. Cependant, dans la réalité ce chemin linéaire entre deux hôtes existe rarement à cause des politiques de routage et des congestions pouvant occasionner un délai supplémentaire dans les mesures.

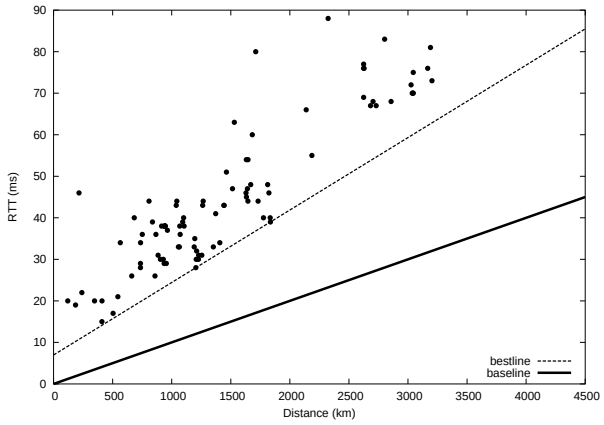


FIG. 3.2 – Exemple montrant la relation entre délai et distance géographique.

Ainsi pour modéliser la relation entre délai et distance géographique, nous utilisons une droite nommée "bestline", que l'on peut définir comme étant une *"meilleure" droite*. L'appelation de la bestline comme la "meilleure" droite est du fait qu'elle essaie de capturer la meilleure relation existant entre délai et distance géographique. La bestline peut être définie comme la droite $y = m_i x + b_i$ qui est la plus proche, mais en dessous de tous les points (x, y) [86] et dont l'ordonnée à l'origine *i*.e. b_i n'est pas négative. Considérer un délai négatif serait un non sens.

Nous formulons l'idée ci-dessus de la bestline comme un problème de programmation linéaire [87]. La condition pour que la bestline soit en dessous de tous les points constitue la première partie de notre problème de programmation linéaire et définit la possible région de notre solution ; la fonction objective de notre problème linéaire minimise la distance entre la droite et l'ensemble de tous les points.

3.3.1 Algorithme de la bestline

La bestline est considérée comme la droite qui prend compte la distorsion la plus petite entre le délai et la distance géographique. L'ordonnée à l'origine de cette droite traduit la présence d'une source de distorsion. Ainsi, la distance entre chaque point et la bestline correspond à la présence d'une source de distorsion par rapport à la bestline qui capture la meilleur relation entre délai et distance. La région qui sépare la bestline à la baseline (*voir* Fig. 3.2) représente l'écart observé entre le cas idéal et la relation existant entre délai et distance géographique à l'intérieur du réseau. Chaque hôte référence L_i calcule sa propre bestline par rapport à tous les autres hôtes références restants.

Soit un hôte référence L_i, nous calculons le délai d_{ij} et la distance géographique g_{ij} vers chaque hôte référence L_j, où $i \neq j$. Nous cherchons alors pour chaque hôte référence L_i la pente m_i et l'ordonnée à l'origine b_i qui déterminent la bestline :

$$y = m_i x + b_i \cdot \tag{3.2}$$

La bestline de chaque hôte référence L_i est en dessous de tous les points (x, y) s'il existe une région formée par l'ensemble des couples (x, y) solution de :

$$y - m_i x - b_i \geq 0, \quad \forall i \neq j, \tag{3.3}$$

La fonction objective qui minimise la distance entre la droite dont l'ordonnée à l'origine est positive et les mesures de délai est définie par

$$\min_{\substack{b_i \geq 0 \\ m_i \geq m}} \left(\sum_{i \neq j} y - m_i x - b_i \right), \tag{3.4}$$

où m représente la pente de la baseline (droite théorique).

L'équation 3.3 détermine la meilleure droite (bestline) de chaque hôte référence H_i. Quant à l'équation 3.4, elle est utilisée pour trouver m_i et b_i. Chaque hôte référence utilise sa propre bestline pour convertir le délai obtenu, entre l'hôte cible et lui, en distance géographique surestimée. Ainsi, la distance géographique surestimée $\hat{g}_{i\tau}$, établie entre l'hôte référence L_i et l'hôte cible τ, est obtenue à partir de l'équation ci-dessous

$$\hat{g}_{i\tau} = \frac{d_{i\tau} - b_i}{m_i}. \tag{3.5}$$

où $d_{i\tau}$, m_i, et b_i représentent respectivement le délai mesuré entre l'hôte référence L_i et la cible τ, la pente, et l'ordonnée à l'origine de la bestline de L_i.

Si les mesures de délai entre les hôtes références se font périodiquement, chaque hôte référence est capable d'ajuster sa propre vision de la relation existant entre le délai et la distance géographique par rapport à l'état du réseau.

3.4 Application de la multilatération

La technique CBG utilise une approche géométrique pour estimer la localisation d'un hôte cible τ. Chaque hôte référence L_i infère la distance géographique surestimée $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$ qui la sépare de l'hôte cible τ en utilisant l'équation 3.5. Ainsi, chaque hôte référence L_i estime que l'hôte cible τ se trouve quelque part à l'intérieur du cercle $\mathcal{C}_{i\tau}$ centré en L_i et de rayon $\hat{g}_{i\tau}$ (analogue à l'exemple de la Fig. 3.1). Par exemple si nous avons K hôtes références, nous avons $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \ldots, \mathcal{C}_{K\tau}\}$ cercles qui constituent un diagramme de Venn [88] d'ordre K avec 2^K régions possibles où on peut localiser l'hôte cible. Cependant on s'intéresse à l'unique région $\mathcal{R}$ formée par l'intersection de l'ensemble des cercles $\mathcal{C}_{i\tau} \in \mathbf{C}_\tau$. Cette région $\mathcal{R}$ est donnée par l'équation

$$\mathcal{R} = \bigcap_{i}^{K} \mathcal{C}_{i\tau}. \tag{3.6}$$

Cette région $\mathcal{R}$ correspond par exemple à la zone grise de la figure (Fig. 3.1) et doit contenir la position réelle de l'hôte cible. La région $\mathcal{R}$ ainsi obtenue est convexe. Ceci est dû au fait que les cercles $\mathcal{C}_{i\tau}$ sont tous convexes, et par définition l'intersection d'un ensemble de convexes est convexe.

La distance géographique obtenue à partir de la transformation du délai est surestimée. Par conséquent, chaque hôte référence surestime la distance géographique réelle entre lui et la cible, obtenant ainsi une région $\mathcal{R}$ formée par l'intersection de l'ensemble des cercles. Si la baseline était utilisée pour transformer les délais en distance, alors la distance géographique obtenue serait fortement surestimée.

Ceci est dû fait que nos mesures de délai sont réalisées dans un milieu non idéal *i.e.* présence de distorsions. Avec une surestimation assez importante de la distance géographique, causée par l'utilisation de la baseline, nous obtenons une vaste région d'intersection $\mathcal{R}$ conduisant à une mauvaise estimation de localisation d'un hôte cible. Quant à la bestline, elle capture la meilleure relation entre délai et distance géographique observée à l'intérieur du réseau. L'idée principale qui est derrière l'utilisation de la bestline est de minimiser la surestimation de la distance géographique en tenant compte de l'état du réseau. L'utilisation d'un certain nombre d'hôtes références tend à apporter de la diversité dans le calcul de la bestline de sorte que la bestline capture le mieux l'état du réseau pour un ensemble de points de références fixés.

3.5 Effets de la surestimation ou de la sous-estimation de la distance géographique sur la multilatération

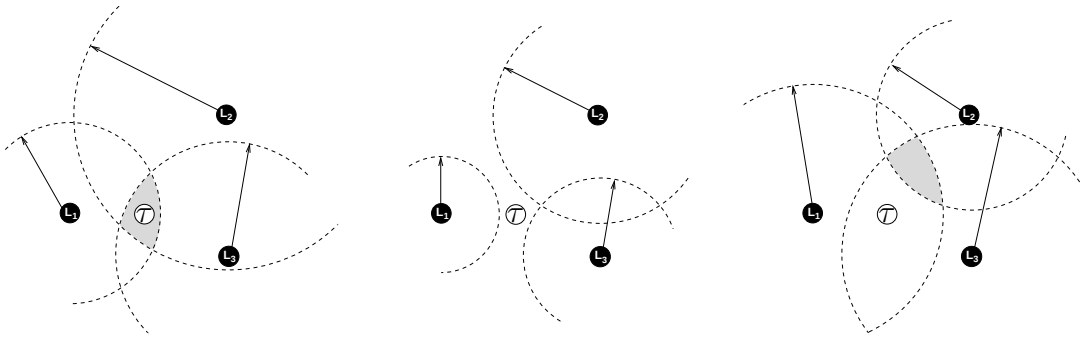
En établissant l'ensemble $\mathbf{C}_\tau$ des cercles pour localiser un hôte cible τ, trois situations peuvent arriver :

- Tous les hôtes références surestiment leur distance géographique (obtenue par la transformation du délai) vers la cible.
- Tous les hôtes références sous-estiment leur distance géographique.
- Certains hôtes références surestiment leur distance géographique et d'autres sous-estiment, conduisant une discordance parmi les hôtes références.

La figure (Fig. 3.3) décrit ces trois situations.

Sur la figure (Fig. 3.3(a)), toutes les distances géographiques obtenues à partir de la transformation du délai sont surestimées. Par conséquent, CBG trouve une zone d'intersection $\mathcal{R}$ qui permet d'inférer l'estimation de localisation de la cible τ. Nous nous attendons à ce que cette situation (obtention d'une zone d'intersection) soit la seule qui puisse arriver, si un nombre suffisant d'hôtes références est utilisé. En effet, les résultats expérimentaux présentés dans le chapitre 4 confirment cette hypothèse pour l'ensemble des hôtes cibles considérés.

Si les distances géographiques obtenues entre l'ensemble des hôtes références et la cible τ sont sous-estimées, comme montré sur la figure (Fig. 3.3(b)), la région $\mathcal{R}$ est vide, *i.e.* il n'y a pas de zone d'intersection. Cette situation arrive seulement, si du point de vue des hôtes références, l'hôte cible présente un ratio, entre délai et distance géographique, plus petit que celui capturé par la bestline de chaque hôte référence. Ceci est improbable. Dans cette situation, en se basant sur l'approche de la bestline, CBG ne peut pas inférer la localisation de la cible. Par conséquent, CBG déclare qu'une estimation de localisation n'est pas possible



(a) Surestimation de la distance géographique (b) Sous-estimation de la distance géographique (c) Discordance

FIG. 3.3 – Effets de la surestimation ou de la sous-estimation de la distance géographique.

pour cet hôte cible τ. C'est une propriété importante de CBG, car pour plusieurs applications, il est préférable de ne pas obtenir d'estimation plutôt de recevoir une mauvaise estimation faite de manière aveugle.

Sur la figure (Fig. 3.3(c)), nous illustrons une situation où deux hôtes références, L_1 et L_3 surestiment leur distance géographique, obtenue à partir des mesures de délai, tandis que l'hôte référence L_2 sous-estime sa distance géographique obtenue. Cette discordance, au niveau des distances géographiques surestimées obtenues, conduit à une zone d'intersection qui ne contient pas l'hôte cible τ. Dans ce cas, la méthodologie CBG tombe en échec, car la position de l'hôte cible va être inférée à l'intérieur de la zone d'intersection obtenue, alors qu'il ne s'y trouve pas en réalité. Nous pensons que cette situation est improbable.

En effet, considérons dans un premier temps deux groupes d'hôtes références : le premier est celui des hôtes références qui surestiment leur distance géographique vers la cible, et l'autre groupe ceux qui sous-estiment leur distance géographique vers la cible. La situation de discordance survient quand la relation observée, entre délai et distance géographique entre ces deux groupes vers l'hôte cible, est assez déséquilibrée. Bien que nous savons que le routage asymétrique dans l'Internet soit assez fréquent (et comme conséquence une capacité asymétrique entre les chemins), nous pensons que la différence de capacité n'est pas assez importante pour être à l'origine de cette situation de discordance entre nos deux groupes d'hôtes références. En outre, la phase d'auto-calibration de la méthode CBG, lors de la conception de chaque bestline, permet à l'ensemble des hôtes références de tenir compte de l'état courant du réseau. Ainsi, chaque hôte référence a un point

de vue unilatéral par rapport aux autres hôtes références restants, et peut donc tenir compte d'éventuelles asymétries notées à l'intérieur du réseau.

3.6 Conclusion

La technique CBG est capable de transformer les mesures de délai en distances géographiques surestimées. La multilatération appliquée à ces distances géographiques surestimées permet d'obtenir une zone d'intersection dans laquelle CBG infère l'estimation de localisation de l'hôte cible. Dans le cas où, les distances géographiques estimées obtenues, entre les hôtes références et la cible, sont sous-estimées, CBG est capable de détecter cette situation et de ne pas fournir d'estimation de localisation. L'auto-calibration que fait CBG permet d'éviter une situation de discordance entre les hôtes références dans laquelle CBG tombe en échec. Nous confirmons que dans toutes nos expérimentations (voir chapitre 4), les distances géographiques obtenues lors de la transformation du délai sont toujours surestimées permettant ainsi d'inférer l'estimation de la localisation géographique de l'hôte cible.

Chapitre 4

Évaluation de CBG

Dans ce chapitre, nous présentons les résultats obtenus lors de l'évaluation de l'approche CBG. Nous montrons également comment CBG infère l'estimation de localisation de l'hôte cible à partir de la zone d'intersection $\mathcal{R}$. Pour ce faire, nous avons considéré des ensembles de données provenant de *NLANR* [89] et de *RIPE* [90] comme paramètres expérimentaux (les mesures de délai étaient déjà disponibles), et notre propre infrastructure de localisation géographique *GeoLIM* [91], déployée sur *PlanetLab* [92].

4.1 Déploiement de GeoLIM sur PlanetLab

Le projet GeoLIM (*Geographic Location of Internet hosts with Multilateration*) est une implémentation de l'approche CBG, et est déployé sur PlanetLab. PlanetLab [92] est un réseau à couches (*overlay*) qui permet de faire des expérimentations à très large échelle et dans un réseau "réel" [93]. C'est une plate-forme géographiquement distribuée, sur l'ensemble des continents exceptés l'Afrique, et formée par quelques 300 sites. Chaque site peut avoir un ou plusieurs nœuds (machines). Même si PlanetLab ne représente pas l'Internet globalement [23], c'est le seul réseau public à grande échelle, où l'on peut faire des mesures à grandeur nature. Ainsi, le choix de PlanetLab pour implémenter l'outil GeoLIM s'est imposé naturellement.

Nous avons implémenté l'outil GeoLIM [91] sur un nœud de PlanetLab qui joue le rôle de serveur. Le système d'exploitation qui tourne sur notre serveur ainsi que sur les hôtes références est *Linux Fedora core 4*. Le langage de programmation et l'environnement utilisé est *R* [94] qui est un logiciel libre, et qui tourne sur la plupart des systèmes d'exploitation (Unix, Windows, Mac OS).

Sur la figure (Fig. 4.1), dés que le serveur reçoit une requête d'un client qui désire se faire localiser (étape 1), il se connecte sur notre ensemble d'hôtes références (nœuds PlanetLab) par *ssh* [95] (étape 2), et demande à ces derniers de mesurer le RTT vers l'hôte cible (étape 3). L'authentification entre le serveur et les hôtes références se fait par clé privée et clé publique. L'ensemble des hôtes références possèdent chacun la clé publique du serveur de localisation. Il faut noter que les mesures de RTT, faites par les hôtes références vers l'hôte cible, se font en parallèle. Pour mesurer le RTT, chaque hôte référence envoie 10 paquets espacés d'une minute chacun, et le RTT minimum est choisi. L'ensemble des hôtes références retourne au serveur le RTT minimum mesuré entre eux et la cible (étape 4). S'il arrive qu'un hôte référence ne retourne pas de RTT, il n'est pas considéré dans le processus de localisation de la cible. Le serveur par le biais de la bestline de chaque hôte référence, transforme les mesures de délai correspondantes entre l'hôte cible et cet hôte référence en distance géographique. Le serveur de localisation fournit ensuite l'estimation de localisation de la cible (étape 5). Le processus de localisation de la cible est expliqué plus en détail dans la section 4.2.2.

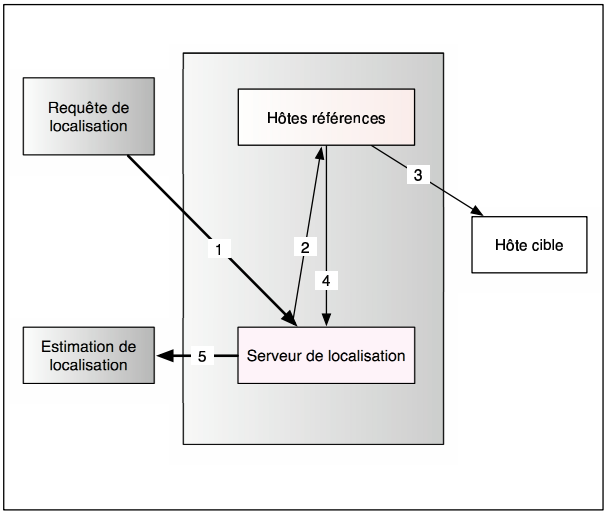


FIG. 4.1 – Architecture de GeoLIM déployée sur PlanetLab.

4.2 Évaluation de CBG avec des ensembles de données

4.2.1 Paramètres expérimentaux

Pour pouvoir évaluer la précision de CBG, nous n'avons besoin que des données fournies par des hôtes dont on connaît leur localisation géographique. Ainsi, pour nos premières expériences nous avons eu en notre possession deux ensembles de données. Ces deux ensembles de données sont :

- RIPE : les données ont été collectées à partir du réseau européen RIPE [90] grace au projet *TTM* (*Test Traffic Measurements*). Nous avons considéré les 2.5 centiles du délai de bout en bout entre 42 hôtes RIPE pendant 10 semaines durant la période de Décembre 2002 à Février 2003. Chaque hôte RIPE, par jour, génère un volume de trafic approximatif de 300 kB vers chaque autre hôte RIPE avec une moyenne de 2 paquets envoyés par minutes. Chaque hôte RIPE est équipé d'une carte GPS permettant ainsi de connaître sa position géographique. Ces 42 hôtes RIPE forment notre ensemble d'hôtes références pour l'Europe de l'Ouest (E.O). La figure (Fig. 4.2(a)) montre la distribution des hôtes RIPE.
- NLANR AMP : les données étudiées ont été collectées grâce à l'organisation NLANR [89] par le biais de son projet *AMP* (*Active Measurement Project*). Les 2.5 centiles du RTT (Round Trip Time) mesuré entre les 95 hôtes localisés à l'intérieur des États Unis sont considérés. Ces mesures de RTT ont été collectées durant la journée du 30 Janvier 2003 et sont symétriques. Les RTT sont prélevés en moyenne une fois par minute, ce qui fait que chaque hôte AMP génère un volume de trafic de 144 kB par jour vers chaque autre hôte AMP. La localisation géographique de chaque hôte (latitude et longitude) est connue. Ces 95 hôtes forment notre ensemble d'hôtes références pour les États Unis. La distribution géographique des hôtes AMP est illustrée sur la figure (Fig. 4.2(b)).

Seul le RTT minimum, mesuré entre les hôtes, est considéré pour l'ensemble de nos deux ensembles de données. Il est plus vraisemblable, qu'il reflète le mieux le délai de propagation et qu'il soit le moins assujetti aux congestions et autres sources de distorsions. Il faut noter que tout RTT correctement mesuré, entre deux nœuds de bout en bout, ne peut être une sous-estimation du RTT minimum. Toutefois, nous considérons en effet que certains RTTs peuvent être mesurés incorrectement. Nous utilisons ainsi les 2.5 centiles du RTT minimum pour éviter les possibles erreurs qui peuvent s'introduire dans les mesures.

Les paramètres expérimentaux considérés comprennent seulement des hôtes situés aux États Unis et en Europe Occidentale. La principale raison de cette

(a) 42 hôtes références situés en Europe de l'Ouest

(b) 95 hôtes références situés aux U.S.

FIG. 4.2 – Distribution géographique des hôtes références (pas à la même échelle).

restriction est due au fait que les données mises à notre disposition contiennent seulement des hôtes localisés dans ces zones. Nous pensons que les résultats obtenus avec ces données sont intéressantes, même si limités aux États Unis et à l'Europe occidentale. Toutefois, dans la section 4.3, nous présentons des résultats obtenus avec des hôtes se trouvant dans d'autres zones géographiques.

Nous construisons deux matrices de délai $\mathbf{D}_{\text{ripe}}$ de dimension (42×42) et $\mathbf{D}_{\text{amp}}$ de dimension (95×95) à partir des mesures obtenues dans chaque ensemble. Chaque hôte étant considéré comme hôte référence (landmark), nous obtenons ainsi deux ensembles d'hôtes références : $\mathcal{L}_{\text{ripe}} = \{L_1, L_2, \ldots, L_{42}\}$ et $\mathcal{L}_{\text{amp}} = \{L_1, L_2, \ldots, L_{95}\}$. Ensuite, comme décrit dans la section 3.3, nous cherchons l'ensemble des bestlines pour chaque élément appartenant à l'ensemble des hôtes références $\mathcal{L}_{\text{ripe}}$ et $\mathcal{L}_{\text{amp}}$. Le calcul de la bestline de chaque hôte référence se fait en considérant seulement les hôtes références appartenant au même ensemble que lui. Par exemple le calcul de la bestline d'un hôte référence appartenant à l'ensemble de données RIPE se fait par rapport aux autres hôtes références de ce même ensemble de données.

L'ensemble des bestlines, de chaque ensemble $\mathcal{L}_{\text{ripe}}$ et $\mathcal{L}_{\text{amp}}$, est défini par les vecteurs $\mathbf{m} = [m_1, m_2, \ldots, m_i]^T$ et $\mathbf{b} = [b_1, b_2, \ldots, b_i]^T$ qui représentent respectivement la pente et l'ordonnée à l'origine de la droite $y = mx + b$ (*i*.e. la bestline).

Après avoir calculé la bestline de chaque hôte référence, dans chaque ensemble, nous transformons les mesures de délai en distances géographiques surestimées en utilisant l'équation 3.5. Nous obtenons ainsi deux matrices $\mathbf{G}_{\text{ripe}}$ et $\mathbf{G}_{\text{amp}}$ formées par les distances géographiques surestimées. Ces matrices contiennent les distances que nous utilisons pour évaluer notre approche.

Dans nos expériences, les hôtes références de chaque ensemble jouent à tour de rôle l'hôte cible à localiser. Les hôtes références restants, appartenant au même ensemble tentent de le localiser. Il faut noter que la bestline de l'hôte référence choisi comme hôte cible n'est pas utilisée lors de sa localisation c'est seulement les bestlines des autres hôtes références restants qui sont utilisées. Ce processus est répété pour la localisation de chaque hôte pris comme cible dans de nos ensembles de données des États Unis et et de l'Europe Occidentale.

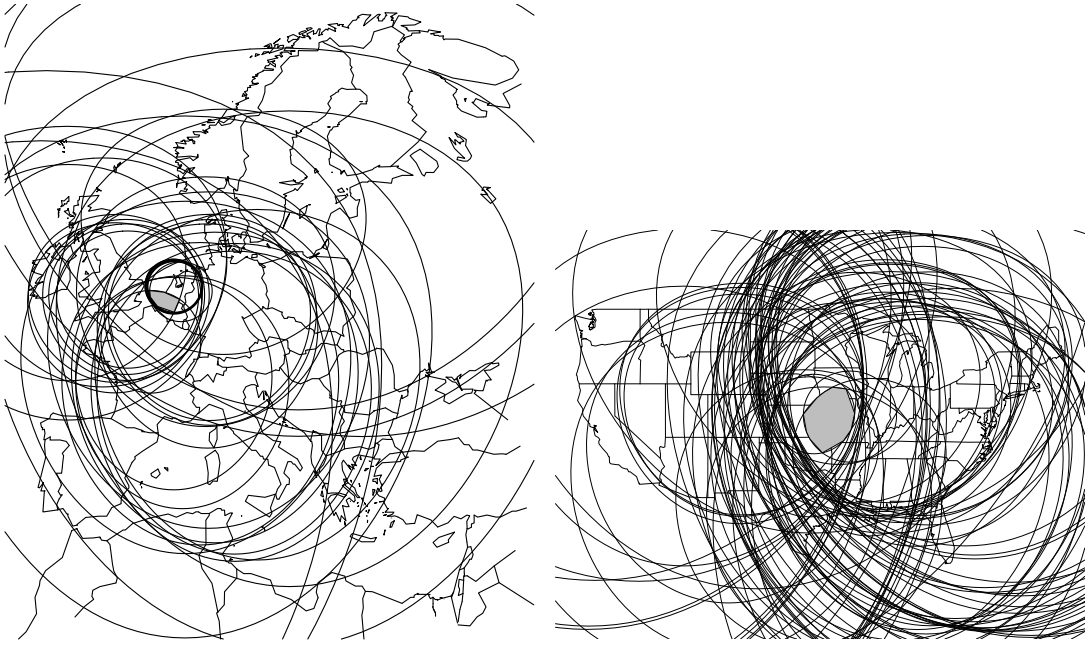
4.2.2 Recherche de la zone géographique d'un hôte cible

A partir des distances géographiques surestimées répertoriées dans les matrices $\mathbf{G}_{\text{ripe}}$ et $\mathbf{G}_{\text{amp}}$, CBG détermine pour chaque hôte cible τ l'ensemble des cercles $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \ldots, \mathcal{C}_{K\tau}\}$ (*v*oir section 3.4) où K=95 pour l'ensemble de données U.S. et K=42 pour celle de l'Europe Occidentale. Chaque cercle de $\mathbf{C}_\tau$ a pour centre la position géographique de son hôte référence L_i correspondant et pour rayon la distance géographique surestimée $\hat{g}_{i\tau}$. La figure (Fig. 4.3) montre un exemple extrait à partir de nos résultats obtenus et illustre la méthodologie de CBG. La figure (Fig. 4.3(a)) illustre l'estimation de localisation d'un hôte RIPE situé à Bruxelles (Belgique). Nous avons 41 cercles qui montrent comment cet hôte cible est vu par l'ensemble des autres hôtes références restants de l'Europe Occidentale. De la même manière, la figure (Fig. 4.3(b)) montre l'ensemble des 94 cercles utilisés pour estimer la localisation d'un hôte AMP situé à Lawrence, en Kansas aux États Unis.

La région grise illustrée au niveau des figures (Fig. 4.3(a)) et (Fig. 4.3(b)) représente la zone d'intersection $\mathcal{R}$ de ces cercles. Cette région $\mathcal{R}$ est la zone de confiance que CBG associe à chaque estimation de localisation. Cependant la plupart des hôtes localisés ont une zone de confiance assez petite. C'est pour mieux illustrer la technique CBG que l'exemple de la figure (Fig. 4.3) a été choisi. La section 4.2.4 répertorie la taille des différentes zones de confiance obtenues.

4.2.3 Processus de localisation d'un hôte cible

La région $\mathcal{R}$ représente l'estimation de localisation fournie par CBG pour inférer la position de l'hôte cible τ. Une idée raisonnable est de prendre le centre de cette région comme la localisation de l'hôte cible. Cependant nous ne savons



(a) Hôte RIPE localisé à Bruxelles, Belgique (b) Hôte AMP localisé en Kansas, U.S.

FIG. 4.3 – Exemples de localisation d'un hôte cible (pas à la même échelle).

pas déterminer la forme géométrique correspondante à cette région, d'où trouver son centre est impossible.

Ainsi, pour estimer la région $\mathcal{R}$, nous choisissons comme heuristique le polygone qui y est inscrit. Le polygone obtenu fournit l'estimation de localisation de l'hôte cible, et sa surface la zone de confiance associée à l'estimation de localisation. Pour former ce polygone, nous considérons les points d'intersection des cercles $\mathcal{C}_{i\tau}$, et se trouvant à l'intérieur de tous les cercles $\mathcal{C}_{i\tau}$, comme sommets du polygone. Le polygone obtenu représente une sous-estimation de la région $\mathcal{R}$. Ceci est du au fait que la région $\mathcal{R}$ est une surface convexe. Par exemple, sur la figure (Fig. 3.1), les sommets de notre polygone vont être les points où se croisent les lignes en pointillées et appartenant à la région colorée en grise. Ainsi chaque polygone est formé par des segments de droites qui relient les N sommets $v_n = (x_n, y_n)$, $0 \leq n \leq N-1$ entre eux. Le dernier sommet du polygone $v_N = (x_N, y_N)$ est supposé être le premier, *i.e.* le polygone est fermé. La surface d'un polygone convexe avec pour sommets $v_0 = (x_0, y_0), \ldots, v_{N-1} = (x_{N-1}, y_{N-1})$ est donnée par

$$A = \frac{1}{2} \sum_{n=0}^{N-1} \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \tag{4.1}$$

(a) Hôte RIPE situés à Bruxelles, Belgique (b) Hôte AMP situés à Kansas, U.S.

FIG. 4.4 – Estimation de localisation à partir de l'heuristique du polygone (pas à la même échelle).

où $|\mathbf{M}|$ représente le déterminant de la matrice $\mathbf{M}$. Le centre du polygone c, *i.e.* l'endroit où est estimé la position de l'hôte cible τ, à pour coordonnées le couple (c_x, c_y) égal à

$$c_x = \frac{1}{6A} \sum_{n=0}^{N-1} (x_n + x_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \tag{4.2}$$

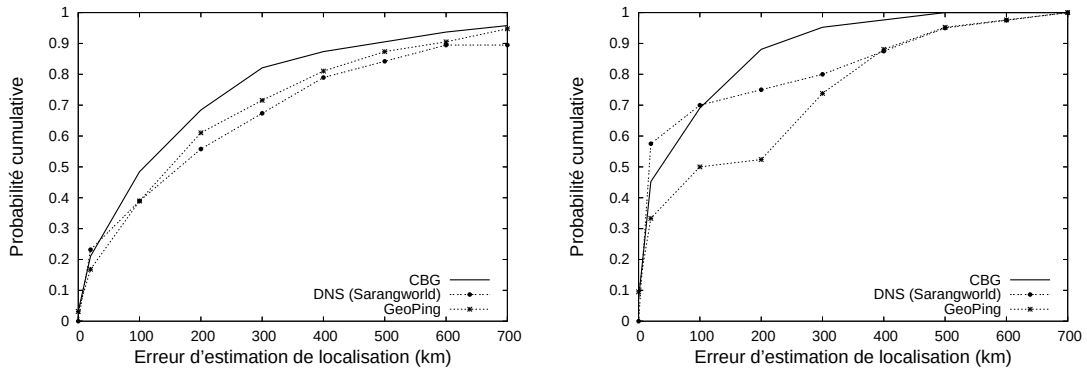
et

$$c_y = \frac{1}{6A} \sum_{n=0}^{N-1} (y_n + y_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix}. \tag{4.3}$$

Le point de coordonnées (c_x, c_y) et la surface $\mathcal{A}$ représentent respectivement l'estimation de localisation de l'hôte cible et la zone de confiance associée à cette estimation.

La figure (Fig. 4.4) illustre un exemple de localisation de deux hôtes à grâce à l'heuristique du polygone. La surface grise illustrée sur la figure (Fig. 4.4) représente la surface du polygone obtenue après approximation de la zone d'intersection des cercles 4.3. La position réelle de l'hôte cible est représentée par un point tandis que l'estimation de localisation de l'hôte cible, donnée par le centre du polygone, est représentée par la croix (*v*oir Fig. 4.4).

Après avoir trouvé la position d'estimation de chaque hôte cible, nous avons calculé l'erreur de distance qui représente la différence entre la position estimée et la position réelle de l'hôte cible τ. Nous avons comparé nos résultats avec ceux obtenus par des méthodes de localisation géographique basées sur les noms DNS et basées sur des mesures de délai (GeoPing). La méthode basée sur le nom DNS (*voir* le projet SarangWorld Traceroute [55]), fait des *traceroutes* [56] vers l'hôte cible et infère la position des routeurs intermédiaires à partir de leur nom DNS. La position géographique du dernier routeur "reconnaissable" sur le chemin est la localisation géographique de l'hôte cible. Un routeur est dit reconnaissable si sa localisation géographique peut être déduite à partir de son nom DNS. La méthode GeoPing utilise un espace discret de réponses [52], *i.e.* la position géographique de l'ensemble des hôtes références est choisie comme possible estimation de localisation des hôtes cibles.



(a) Hôtes situés aux U.S.

(b) Hôtes situés en Europe Occidentale

FIG. 4.5 – Erreur d'estimation de localisation de CBG, de la méthode DNS et de GeoPing.

La figure (Fig. 4.5) montre la fonction de probabilité cumulative de l'erreur d'estimation de localisation obtenue en utilisant CBG, la méthode basée sur le DNS et GeoPing. CBG dépasse en précision et la méthode basée sur le DNS et la technique GeoPing. De plus, l'écart noté au niveau de l'Europe Occidentale est important. Ceci est dû probablement au fait que le nombre d'hôtes références, qui y est localisé, est moins important que celui des États Unis. Il est montré dans [71] que si nous avons un espace discret d'endroits où on peut localiser un hôte, le nombre d'hôtes références et leur placement jouent un rôle considérable dans la précision de l'estimation de localisation. Dans la section 4.2.5, nous montrons l'impact du nombre d'hôtes références sur les performances de CBG.

Sur la figure (Fig. 4.6), nous comparons les erreurs de distance obtenues lors de la localisation des hôtes situés aux U.S. et en Europe Occidentale en utilisant l'approche CBG. L'erreur moyenne d'estimation de localisation est évaluée à

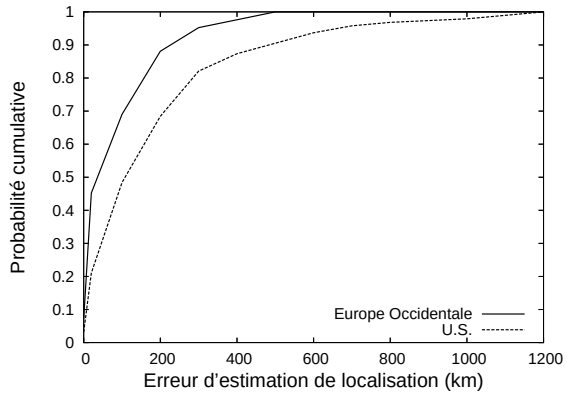


FIG. 4.6 – Erreur d'estimation de localisation de CBG pour les ensembles de données U.S. et de l'Europe Occidentale

182 km au niveau des U.S. alors qu'elle est de 78 km pour l'Europe Occidentale. La plupart des hôtes ont une assez bonne estimation de localisation au niveau de nos deux ensembles de données. L'erreur médiane de l'estimation de localisation est de 95 km et 80 % des hôtes sont localisés avec une erreur inférieure à 277 km. Pour l'ensemble des hôtes situés en Europe Occidentale, l'erreur médiane est de 22 km et 80 % sont localisés avec une erreur inférieure à 134 km. Nous identifions et décrivons les possibles raisons de l'imprécision des mesures plus en détail dans la section 4.2.6.

4.2.4 Zone de confiance associée à l'estimation de localisation

La surface totale de la zone d'intersection $\mathcal{R}$ est assez liée à la zone de confiance que CBG associe à chaque estimation de localisation. Intuitivement, cette zone représente l'étendue de chaque estimation de localisation évaluée en km^2. Plus la surface $\mathcal{R}$ est petite, plus la confiance que nous pouvons avoir sur l'estimation de localisation fournie par CBG est élevé. Ainsi, la force de CBG est sa capacité à fournir une zone de confiance, contrairement aux techniques précédentes de localisation. Cette zone de confiance est importante pour les applications qui l'utilisent afin d'évaluer le niveau de fiabilité par rapport à leurs exigences.

La figure (Fig. 4.7) illustre la fonction de probabilité cumulative des zones de confiance associées à l'estimation de localisation des hôtes références situés aux U.S. et en Europe Occidentale. Les résultats montrent que pour les U.S., 80 % des hôtes ont la surface de leur zone de confiance inférieure à 10^5 km^2. Par exemple cette surface est légèrement supérieure à la superficie d'un pays comme le Portugal ou d'un état des États Unis comme l'Indiana. Pour les hôtes situés

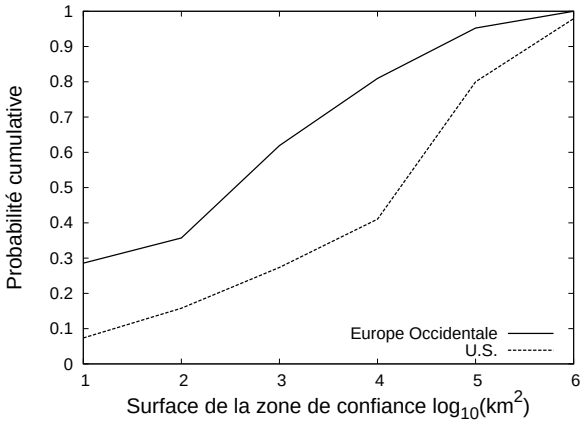


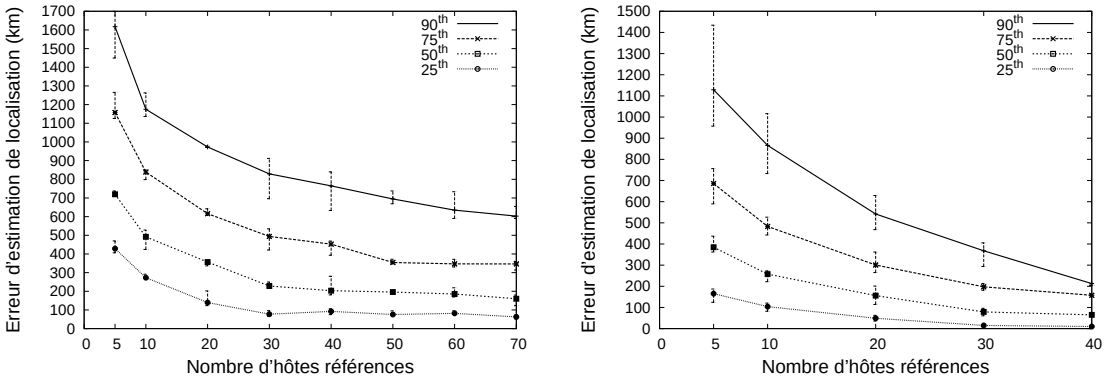
FIG. 4.7 – Zone de confiance fournie par CBG en km^2.

en Europe Occidentale, 80 % ont la surface de leur zone de confiance inférieure à 10^4 km^2 permettant ainsi une localisation à une échelle régionale. En outre, 25 % des hôtes situés aux U.S. sont localisés avec une zone de confiance inférieure à 10^3 km^2 alors qu'en Europe Occidentale 65 % des hôtes le sont. Cette surface est équivalente à la taille d'une grande région métropolitaine.

4.2.5 Impact du nombre d'hôtes références sur la localisation

Dans cette section, nous évaluons l'impact du nombre d'hôtes références sur les performances de CBG. Pour chaque ensemble de données, pour chaque k hôtes références considérés, nous calculons l'erreur moyenne d'estimation de localisation de nos échantillons de mesures. Pour chaque k hôtes références choisis, nous avons considérons 30 échantillons. Ces k hôtes références sont choisis parmi les 42 et 95 hôtes références de nos ensembles de données de l'Europe Occidentale et des U.S. respectivement. Toutefois, comme le nombre de possibilité de placement des hôtes références devient de plus en plus important lorsque k augmente, nous n'avons pas considéré toutes les façons de choisir k hôtes références dans chaque ensemble de données.

La figure (Fig. 4.8) montre différents centiles de l'erreur d'estimation de localisation en fonction du nombre d'hôtes références considérés en utilisant la technique CBG. Par exemple, la courbe qui montre les 90ème de centiles, représente l'erreur d'estimation de localisation, où la courbe de la fonction de probabilité cumulative de l'erreur moyenne d'estimation de localisation rencontre le point dont la probabilité est 90%. Les barres sur la figure (Fig. 4.8) représente nos intervalles de confiance. Ces résultats obtenus montrent qu'un nombre d'hôtes références, en



(a) Ensemble de données U.S. (b) Ensemble de données Europe Occidentale

FIG. 4.8 – Erreur d'estimation de localisation en fonction du nombre d'hôtes références.

l'occurrence 30, est suffisant pour stabiliser l'erreur d'estimation de localisation au niveau de nos deux échantillons de mesures.

4.2.6 Limitations des mesures actives

Deux aspects contribuent à ajouter une robustesse de base à l'estimation de localisation, faite par CBG, en luttant contre les facteurs qui peuvent affaiblir la relation existant entre délai et distance géographique. Premièrement, le délai est mesuré à partir de plusieurs hôtes références dispersés géographiquement, plutôt qu'à partir de trois hôtes références, qui peuvent être suffisants pour appliquer la multilatération comme le fait GPS avec trois satellites. Deuxièmement, parmi plusieurs échantillons de mesures, seul le RTT minimum est considéré afin d'éviter les distorsions dues aux congestions. Au delà de ces deux sources de distorsions, la transformation du délai en distance géographique peut être déformée par d'autres sources qui sont étudiées ci-dessous plus en détails.

4.2.6.1 Non linéarité des chemins entre deux hôtes

Subramanian *et al.* dans [24], en considérant les distances géographiques entre 15 sources et 6000 destinations, étudient comment les routes dans l'Internet sont non linéaires ("*circuitousness*"). Ils montrent également que le degré de connectivité et les politiques de routages entre les systèmes autonomes (AS) sont les causes de la non linéarité des chemins. De plus, au niveau topologie, les routes dans l'Internet ne sont pas nécessairement optimales, puisqu'une route entre deux nœuds terminaux peut être plus longue par rapport à leur distance réelle. Ce phénomène a été analysé tout récemment sous différents noms, tels que "*routing stretch*" [96]

ou "*path inflation*" [97], et contribue à la non linéarité des chemins. Spring et *al.* dans [97] montrent comment le routage intra-domaine, les politiques de *peering* entre les FAI, et le routage inter-domaine contribuent à la non linéarité des chemins (path inflation) entre les hôtes Internet.

Il est aussi bien connu que le routage inter-domaine peut ne pas choisir les meilleures routes disponibles [98, 99, 100]. Le choix du chemin est fortement influencé par les politiques des FAI source, intermédiaire et destinataire. La non linéarité des chemins entre ajoute un délai supplémentaire dans les mesures de délai occasionnant une forte surestimation de la distance géographique estimée.

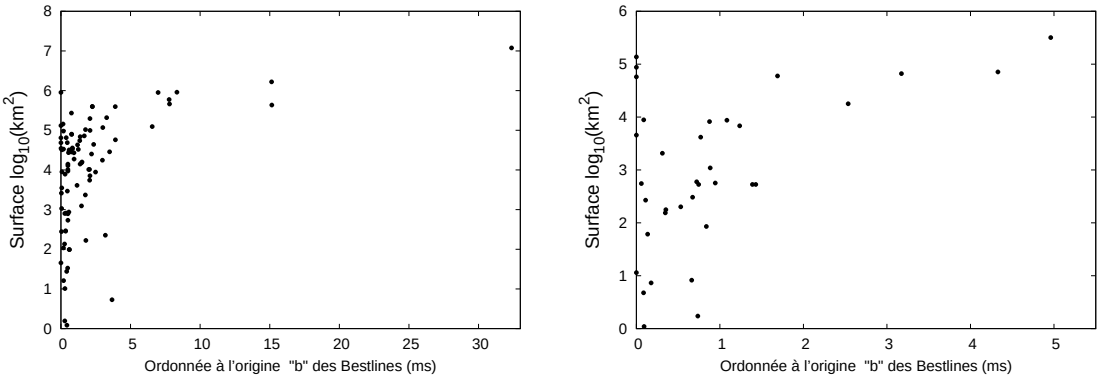
La méthodologie de CBG tente de gérer cette non linéarité existant entre les chemins. Dans le calcul de leur bestline, chaque hôte référence à sa propre vision (*i.e.* auto-calibration) de la relation entre délai et distance géographique. La bestline de chaque hôte référence reflète le chemin qui est le plus proche du chemin théorique (chemin linéaire) représenté par la baseline. Ainsi, chaque bestline tient compte des déviations, par rapport au chemin théorique, notées par l'ensemble des autres hôtes références.

4.2.6.2 Présence de "*localized delay*"

Le phénomène de "*localized delay*" peut être défini comme étant une situation dans laquelle un délai constant s'ajoute au délai mesuré pour un hôte donné. Le localized delay peut advenir dans le cas où on est en présence de liens à faible débit, d'une congestion (goulot d'étranglement), ou bien les deux. Dans notre approche CBG, le localized delay est représenté par l'ordonnée à l'origine b_i au niveau du calcul des bestlines. La présence d'un important localized delay est "trompeur", car elle amène à trop surestimer la distance géographique, obtenue lors de la transformation du délai, menant ainsi à de large zone de confiance.

La figure (Fig. 4.9) compare l'ordonnée à l'origine b_i, de la bestline de chaque hôte référence, et la zone de confiance associée à l'estimation de localisation de cet hôte référence quand il est choisi comme cible. Il faut noter que les figures (Fig. 4.9(a)) et (Fig. 4.9(b)) ne sont pas à la même échelle. Certains hôtes références, situés aux U.S., ont leur bestline définit par un grand b_i conduisant ainsi à de large zone de confiance, contrairement aux hôtes localisés en Europe Occidentale. En outre, quelque soit l'ensemble de données considéré, tous les hôtes références ayant un grand b_i ont une large zone de confiance associée à leur estimation de localisation lorsqu'ils sont choisis comme cible. Cela stipule qu'un important localized delay mène à une large zone de confiance. Toutefois, le contraire n'est pas nécessairement vrai.

La figure (Fig. 4.9) illustre aussi que de petits b_i ne conduisent pas à l'obtention de petites de zone de confiance. Une large zone de confiance peut être une conséquence d'une forte surestimation, par les hôtes références de la "distance



(a) Ensemble de données U.S. (b) Ensemble de données Europe Occidentale

FIG. 4.9 – Zone de confiance en fonction de l'ordonnée à l'origine b ("localized delay").

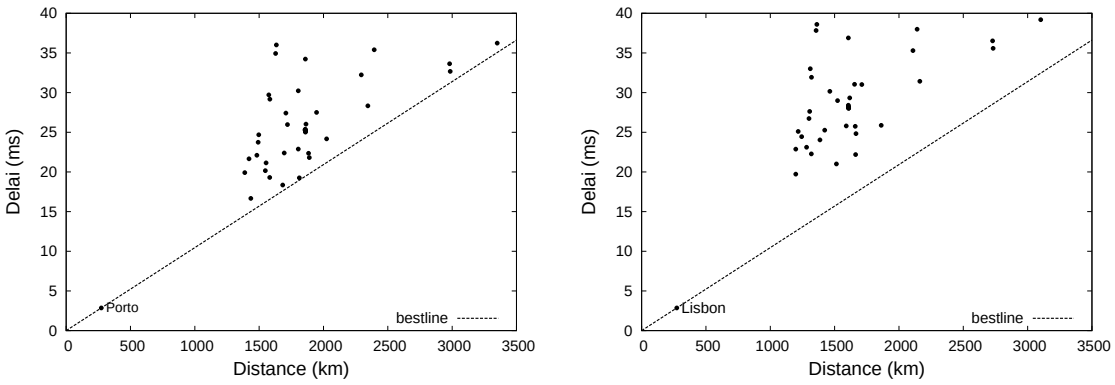
géographique surestimée" vers la cible. Cette forte surestimation n'est pas seulement liée aux conditions locales de la cible, mais elle peut être due à la vision de l'état courant du réseau par les autres hôtes références. Si l'hôte cible se trouve caché dernier un nœud, à cause des chemins partagés ("*shared paths*"), tous les hôtes références vont surestimer la distance géographique estimée même si l'hôte cible ne présente pas de localized delay comme montré dans la section suivante.

4.2.6.3 Chemins partagés ("*shared paths*")

Les mesures de délai faites par les différents hôtes références vers la cible fournissent une information redondante si elles partagent les mêmes chemins. Ainsi, si toutes les mesures passent par un même nœud et partagent l'ensemble des liens restant sur le chemin vers la cible, alors l'estimation de localisation se trouve évaluée aux environs de ce nœud. Ceci conduit à de mauvaises estimations de localisation, *i.e.* de larges zones de confiance. Certains hôtes, comme montré précédemment sur la figure (Fig. 4.9) ont de larges zones de confiance bien qu'ils présentent un faible ou aucun"localized delay". Ces larges zones de confiance sont causées certainement par la présence de chemins partagés ("shared paths") dans nos mesures de délai.

Ce phénomène de chemins partagés est illustré avec l'exemple de deux hôtes RIPE dont l'un est localisé à Lisbonne et l'autre à Porto, deux villes situées au Portugal. Quand l'hôte référence localisé à Porto est choisi comme cible, la zone de confiance associée à son estimation de localisation est égale environ à 57 000 km^2, qui correspond au 2/3 de la superficie du Portugal, d'où une imprécision dans l'estimation de localisation.

La figure (Fig. 4.10) montre comment les bestlines, des hôtes références situés



(a) Bestline de l'hôte référence localisé à Lisbonne

(b) Bestline de l'hôte référence localisé à Porto

FIG. 4.10 – Présence de chemins partagés ("shared paths").

à Porto et à Lisbonne, capturent la relation entre délai et distance géographique à l'intérieur du réseau. Il faut noter que l'hôte référence localisé à Porto détermine la bestline de l'hôte référence localisé à Lisbonne comme illustré sur la figure (Fig. 4.10(a)), et *vice versa* sur la figure (Fig. 4.10(b)). Sur la figure (Fig. 4.10(b)), nous remarquons que, sans la présence de l'hôte référence localisé à Lisbonne, la bestline de l'hôte référence localisé à Porto serait décalée vers les autres hôtes références restants. La figure ainsi obtenue serait certainement semblable à celle de la bestline de l'hôte référence localisé à Lisbonne (Fig. 4.10(a)), sauf qu'une valeur d'environ 5 ms s'ajouterait à l'ordonnée à l'origine b_i de la "nouvelle" bestline de l'hôte référence localisé à Porto.

En effet, le délai mesuré entre l'hôte référence situé à Porto et celui situé à Lisbonne est égal environ à 5 ms. Ainsi, la vision qu'ont les hôtes références restants de l'hôte référence situé à Porto, est la même que celle qu'ils ont vers l'hôte référence situé à Lisbonne plus un délai additionnel de 5 ms. Du point de vue des hôtes références restants, l'hôte référence situé à Porto est dans une certaine mesure caché par l'hôte référence situé à Lisbonne. Ainsi, nous insinuons que tout le trafic à partir de l'hôte référence situé à Porto vers l'ensemble des hôtes références restants, et vice versa, passent par l'hôte référence situé dans la ville de Lisbonne. Comme conséquence, quand l'hôte référence situé à Porto est choisi comme cible, la zone de confiance associée à son estimation de localisation, ressemble à un grand cercle centré aux alentours de Lisbonne. Nous obtenons par la même occasion une mauvaise estimation de localisation.

Nous observons également des cas similaires de chemins partagés au niveau de notre ensemble de données U.S. conduisant à de mauvaises estimations de localisation. Les hôtes AMP `amp-wsu` et `amp-montana`, respectivement situés à Pullman (Washington – WA) et à Bozeman (Montana – MT), semblent être cachés par

l'hôte `amp-uwashington` situé à Seattle (WA). Ainsi, tous les hôtes références restants de notre ensemble de données, par rapport à l'hôte `amp-uwashington`, voient un délai additionnel constant de 10 ms et 15 ms vers les hôtes `amp-wsu` et `amp-montana`. Cette vision mène à l'obtention de larges zones de confiance. Toutes les mesures de délai des autres références vers les hôtes `amp-wsu` et `amp-montana` partagent les mêmes chemins après avoir traversé la ville de Seattle comme le montre les `traceroute` respectifs des traces AMP [89]. Ainsi, le trafic de tous ces hôtes traversent la zone de Seattle.

4.3 Résultats obtenus avec GeoLIM

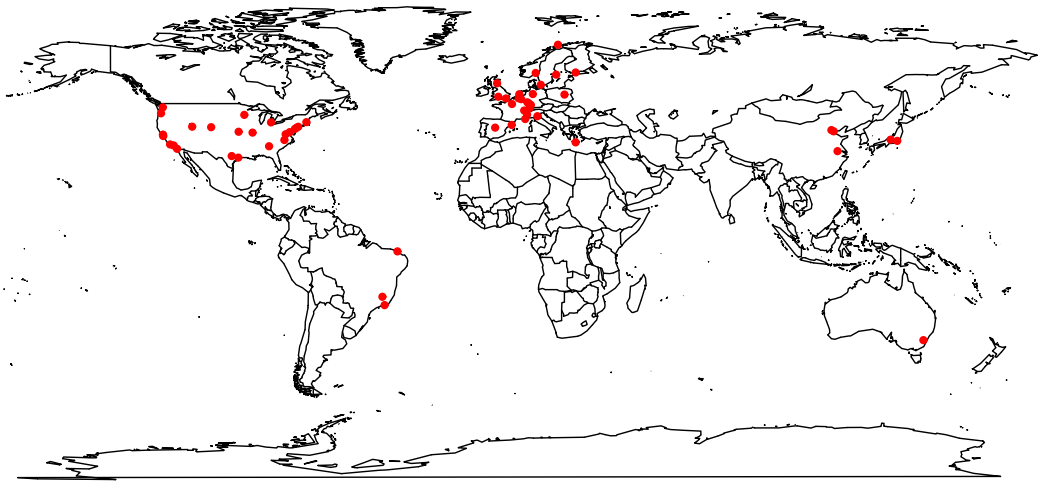


FIG. 4.11 – Distribution géographique des hôte références utilisés par GeoLIM.

Nous avons évalué l'approche CBG sur PlanetLab [92] grâce au projet GeoLIM [91]. L'évaluation a été faite durant le mois de Mai 2005 et 57 hôtes références furent considérés. Les points illustrés sur la figure (Fig. 4.11) montrent la distribution géographique de l'ensemble des hôtes références utilisé par GeoLIM. Ces hôtes références sont des nœuds appartenant au réseau PlanetLab et sont répartis géographiquement comme suit : 24 aux États Unis, 24 en Europe, 5 en Asie, 3 en Amérique du Sud, et 1 en Océanie.

Ces hôtes références ont été utilisés pour la localisation de 42 et 43 hôtes cibles situés respectivement aux U.S. et en Europe, en utilisant la méthodologie CBG. Parmi ces hôtes cibles, nous avons eu 3 qui se sont connectés par modem, 3 par *wifi*, 6 par accès *ADSL*, et le reste s'est connecté par des liens Internet à large bande. Il faut noter que ces hôtes cibles sont des hôtes anonymes qui ont testé

notre plate-forme GeoLIM par son interface Web. Le temps de réponse de chaque estimation de localisation est situé environ entre 2 et 3 minutes.

La figure (Fig. 4.12) montre la probabilité cumulative de l'erreur d'estimation de localisation obtenue lors de l'évaluation de CBG sur PlanetLab, *i.e.* GeoLIM. L'erreur moyenne d'estimation de localisation pour les hôtes cibles situés aux U.S. est de 209 km, tandis que pour les hôtes cibles situés en Europe, elle est de 106 km. L'erreur médiane et l'erreur d'estimation de localisation des 80 centiles des hôtes situés aux U.S. sont égales respectivement à 130 et 411 km. Pour les hôtes cibles situés en Europe, l'erreur médiane d'estimation de localisation est égale à 42 km et les 80 centiles des hôtes sont localisés avec erreur inférieure à 218 km.

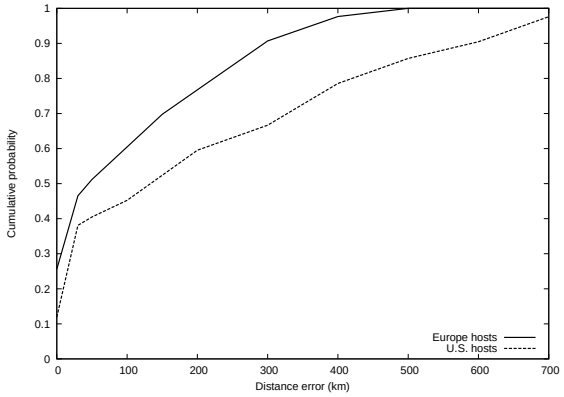


FIG. 4.12 – Erreur d'estimation de localisation avec GeoLIM.

La figure (Fig. 4.13) quant à elle, montre la probabilité cumulative des zones de confiance, en km^2, associées à chaque estimation de localisation fournie par GeoLIM. Nous remarquons que les zones de confiance obtenues avec GeoLIM sont beaucoup plus large que celle obtenue par l'évaluation de CBG avec des ensembles de données (section 4.2). Ceci est certainement dû aux faits que les nœuds RIPE et AMP possèdent une meilleure connectivité que les nœuds anonymes considérés lors de l'évaluation de GeoLIM. Toutefois, nous avons de bonne estimation de localisation avec des zones de confiance inférieures à 10 km^2.

4.4 Conclusion

Transformer les mesures de délai en distance géographique avec précision est un défi en raison de beaucoup de particularités inhérentes à l'Internet. La non linéarité du chemin entre deux hôtes de bout en bout, la présence de "localized delay", et la congestion dans les réseaux pouvant occasionner des délais supplémentaires dans les mesures, sont des exemples des difficultés rencontrées.

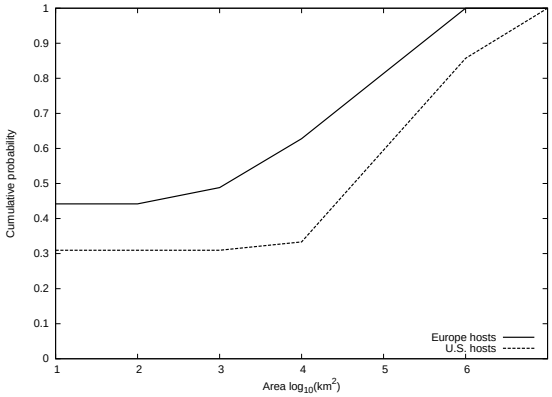


FIG. 4.13 – Zone de confiance en km^2 obtenue avec GeoLIM.

Toutefois, CBG montre qu'il est possible de transformer les mesures de délai en *distances géographiques surestimées* avec des contraintes. La distance géographique obtenue est surestimée assez souvent de manière très petite permettant ainsi de localiser avec précision l'hôte cible en appliquant la multilatération. CBG établit une relation dynamique entre délai et distance géographique grâce à un auto-calibrage des hôtes références par le biais de leur "bestline" respectif.

Les résultats obtenus montrent que la technique CBG est plus performante que les techniques précédentes de localisation géographique. Ainsi, l'erreur médiane de distance obtenue pour les hôtes références localisés aux U.S. est inférieure à 100 km tandis que pour l'Europe Occidentale elle est inférieure à 25 km. Alors que pour la technique GeoPing cette erreur est de 150 km pour les U.S. et 100 km pour l'Europe Occidentale. En outre, contrairement aux méthodes précédentes, CBG associe à chaque estimation de localisation une zone de confiance. Cette zone de confiance est importante, car elle permet aux applications d'évaluer si la précision fournie est suffisante. La plupart des hôtes ont été localisés avec une zone de confiance assez petite de l'ordre d'une région métropolitaine, d'où une précision de notre estimation de localisation.

Nous avons vu que la relation entre délai et distance géographique pouvait être perturbée par différentes sources de distorsions. Ces sources peuvent être la présence de goulets d'étranglements, de "localized delay", la non linéarité des chemins pouvant occasionner de larges zones de confiance, qui conduisent à une imprécision de l'estimation de localisation. Dans le chapitre suivant, nous proposons une technique qui prévoit d'évaluer et de supprimer ces différentes distorsions qui ajoutent un délai supplémentaire au délai mesuré, avant de transformer les mesures de délai en distance géographique surestimée.

Chapitre 5

Influence du délai de buffering sur la localisation

Les techniques de *géolocalisation* basées sur des mesures souffrent des distorsions que subissent les mesures de délai. Ces sources de distorsions peuvent être aussi dues, par exemple, à la congestion qui survient dans les réseaux et à la non linéarité des chemins, ajoutant ainsi un délai supplémentaire dans les mesures de délai [101, 24, 102]. Cette non linéarité des chemins entre les hôtes dans l'Internet peut être due aux politiques de sélection des chemins à l'intérieur des FAI et/ou des politiques de transit entre les FAI voisins [97]. Le temps de traitement des paquets (*buffering*) au niveau des routeurs intermédiaires sur un chemin (source, destination) peut être aussi un facteur de distorsion. Dans ce chapitre, nous examinons l'impact du délai de buffering dans les mesures de délai ainsi que les solutions envisagées pour y remédier.

5.1 Les routeurs : possible source de distorsions

L'Internet est composé d'un ensemble de routeurs et de liaisons de transmission. Les liaisons possèdent des caractéristiques de délai, débit maximum et disponibilité. Les routeurs peuvent avoir un impact significatif sur :

- Le délai : temps écoulé entre l'envoi d'un paquet par un émetteur et sa réception par le destinataire. Le délai tient compte du délai de propagation le long du chemin et du délai de transmission induit par la mise en file d'attente des paquets dans les systèmes intermédiaires.
- La gigue : variation du délai de bout en bout
- La disponibilité : taux moyen d'erreurs d'une liaison

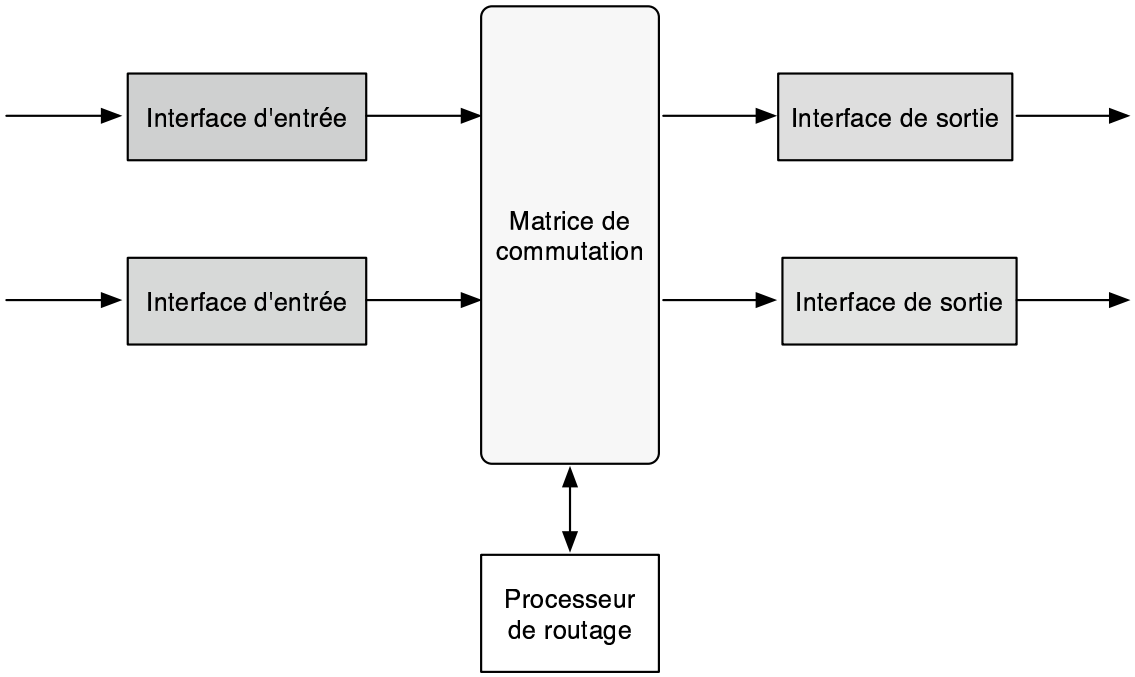


FIG. 5.1 – Architecture d'un routeur (haut de gamme).

La figure (Fig. 5.1) illustre l'architecture d'un routeur. En effet, les fonctions d'un routeur consistent à contrôler l'intégrité du paquet reçu sur l'interface d'entrée, déterminer l'interface de sortie en fonction de la destination souhaitée, puis stocker ce paquet dans la file d'attente associée à celle-ci. Ainsi, la matrice de commutation de la figure (Fig. 5.1) relie les ports entre eux et le processeur de routage se charge de l'exécution des protocoles de routages et de la mise-à-jour des tables de routage. Les interfaces d'entrées contiennent un tampon où sont stockés les paquets arrivant trop vite que la matrice de commutation peut traiter. Ils contiennent également une copie de la table de routage du processeur de routage. Dans chaque interface de sortie, un gestionnaire décide quel paquet de la file doit être transmis. Des politiques d'ordonnancement comme : FIFO(*First In First Out*), le premier paquet qui arrive est le premier qui sort ; WFQ (*Weighted Fair Queuing*) [103], partage équitable entre les flux ; et beaucoup d'autres [104, 105] peuvent être appliquées.

Lorsque le fonctionnement se dégrade (trop de trafic compte tenu des capacités du routeur) les files d'attente se remplissent, introduisant un délai supplémentaire dans la transmission des paquets, puis le routeur est forcé de jeter des paquets. Ces différentes actions ont un impact sur le délai, la gigue et en augmentant la probabilité de paquets transmis dans le désordre, influent également sur la disponibilité.

Ainsi, pour accroître la précision des techniques de géolocalisation basées sur des mesures de délai, il est important d'estimer et de supprimer si possible ces distorsions.

5.2 Introduction à GeoBuD

L'approche GeoBuD ("***Geo****location using* ***Bu****ffering* ***D****elay estimation*") est proposée pour tenir compte de l'influence du temps de traitement (buffering) au niveau des routeurs qui composent le chemin entre une source et une destination. Dans l'Internet, l'outil traceroute [56] permet heureusement de découvrir le chemin entre deux hôtes. Ainsi, nous employons l'outil traceroute pour estimer le temps de traitement introduit au niveau des délais mesurés entre les hôtes références et l'hôte cible. En considérant les différents RTT (Round Trip Time) mesurés au niveau des sauts intermédiaires découverts grâce à l'outil traceroute, nous estimons le délai de buffering introduit au niveau de chaque nœud composant le chemin.

Rechercher l'influence du délai de buffering, induit par chaque routeur, sur la localisation est un véritable défi. En effet, pour estimer avec précision le temps de traitement au niveau des sauts intermédiaires qui composent le chemin suivi par un traceroute, il faut localiser avec exactitude les routeurs intermédiaires. En outre, l'estimation du délai de buffering reste difficile même avec la connaissance de la position géographique des routeurs. Ceci est dû au fait que les mesures de délai contiennent une information grossière, *i.e*, elles peuvent être assujetties à différentes sources de distorsions.

5.3 Méthodologie de GeoBuD

Contrairement à CBG qui se base sur un calibrage entre les hôtes références pour transformer les mesures de délai en distances géographiques (chapitre 3), GeoBuD se base sur le chemin entre chaque hôte référence et l'hôte cible. En effet, dans la pratique, le délai vers différentes destinations peut être assujetti à différentes sources de distorsions. Pour tenir compte de ces contraintes, nous remplaçons le modèle linéaire de CBG par un modèle tenant compte des différents sauts composants le chemin. Ainsi, dans l'approche GeoBuD, pour chaque hôte référence L_i et chaque hôte cible τ nous modélisons le délai $y_{i\tau}$ par

$$y_{i\tau} = m_i x_{i\tau} + b_{i\tau}, \tag{5.1}$$

où m_i représente la pente de la bestline, x_{i_τ} représente la *distance géographique estimée* entre l'hôte référence L_i et la cible τ et $b_{i\tau}$ représente le délai de buffering

estimé sur le chemin entre L_i et l'hôte cible τ. Estimer le délai de buffering revient à l'estimer au niveau de chaque nœud composant le chemin entre l'hôte référence L_i et et l'hôte cible τ. Pour déterminer ce délai de buffering sur le chemin, nous utilisons l'outil traceroute [56].

5.3.1 Traceroute

Le traceroute est utilisé pour connaître la route entre deux stations :
- l'accessibilité
- la liste des routeurs intermédiaires sur cette route
- la durée de transit entre chaque routeur de cette route

Ainsi, le traceroute est capable de fournir le délai (RTT) entre 2 nœuds (routeurs) adjacents consécutifs sur un chemin entre une source et une destination. Toutefois, il faut que ces routeurs intermédiaires répondent par un ICMP (*Internet Control Message Protocol*) [106] `TIME exceeded` pour pouvoir obtenir le RTT. Le protocole ICMP est un protocole de gestion des erreurs lors de la transmission d'un datagramme IP.

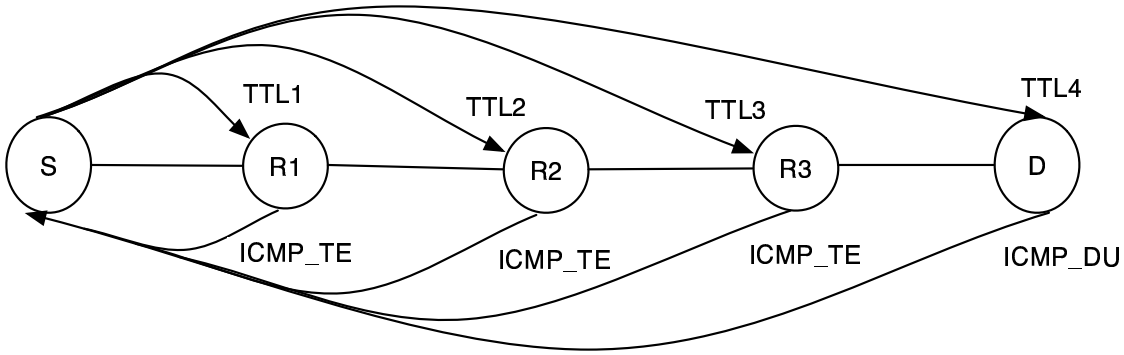


FIG. 5.2 – Exemple de fonctionnement d'un traceroute.

La figure (Fig. 5.2) illustre comment un traceroute fonctionne. Pour se fixer les idées, S est la source du traceroute, D la destination du traceroute et R représentent les routeurs sur le chemin. La source S envoie des paquets UDP (*User Datagram Protocol*) avec un TTL de plus en plus grand (en commençant par 1). Il faut noter que certaines versions de traceroute peuvent aussi utiliser TCP ou bien ICMP avec des paquets `ECHO Request`. Dans nos expérimentations nous avons considéré le traceroute natif [56].

Chaque routeur recevant un paquet IP en décrémente le TTL. Lorsque le TTL atteint 0, le routeur émet un paquet ICMP d'erreur. Traceroute découvre ainsi les routeurs de proche en proche. Le traceroute exécuté sur la figure (Fig. 5.2) est composé de 4 sauts intermédiaires où le dernier saut représente la destination.

Ainsi, pour chaque saut k qui répond par un message ICMP `TIME exceeded` (`ICMP_TE` sur la figure Fig. 5.2), nous obtenons une mesure de RTT. Le destinataire est atteint lorsque la source reçoit un message ICMP `destination unreachable` (`ICMP_DU` sur la figure Fig. 5.2).

Par exemple, la commande *traceroute 132.227.74.40* donne comme résultats pour des *TTL* (*Time To Live*) compris entre 11 et 15 :

11 gw-rap.rap.prd.fr (193.50.20.73) 20.605 ms 20.410 ms 20.435 ms

12 jussieu-rap.rap.prd.fr (195.221.127.182) 20.693 ms 20.804 ms 20.947 ms

13 r-scott.reseau.jussieu.fr (134.157.254.10) 22.094 ms 21.961 ms 22.229 ms

14 olympe-gw.lip6.fr (132.227.109.1) 21.782 ms 21.180 ms 21.193 ms

15 planetlab-01.ipv6.lip6.fr (132.227.74.40) 21.955 ms 22.174 ms 21.839 ms

Malheureusement, un traceroute ne suit pas toujours le comportement décrit sur la figure Fig. 5.2. Un routeur sur le chemin peut ne pas répondre à la source du fait que le protocole ICMP n'y est pas activé ou bien du fait que le routeur est surchargé. Ainsi, a un saut k donné, si un routeur intermédiaire ne répond pas par un message ICMP `TIME exceeded` alors aucun RTT ne lui est associé. Par conséquent, il ne sera pas considéré dans l'estimation du délai de buffering.

5.3.2 Algorithme de GeoBuD

Pour pouvoir estimer le délai de buffering $b_{i\tau}$ dans l'équation (5.1), il faut en fait calculer le temps de traitement b_k au niveau de chaque saut (routeur) (Fig. 5.3) le long du chemin suivi par le traceroute en utilisant

$$\Delta RTT_{k+1} = RTT_{k+1} - RTT_k = m_i \times dist(k, k+1) + b_{k+1}, \qquad (5.2)$$

où k représente le $k^{\text{ème}}$ routeur intermédiaire sur le chemin du traceroute pour lequel nous avons une mesure de délai et dont la localisation géographique est connue. Le terme RTT_k désigne le RTT minimum mesuré à un saut donné k. En effet chaque étape du traceroute consiste à envoyer 3 paquets UDP consécutifs vers une destination en augmentant la valeur du TTL. Dans nos mesures, nous avons utilisé le traceroute natif [56]. Le terme *dist(k,k+1)* représente la distance géographique entre le nœud k et $k+1$. Il faut noter que m_i est le même que celui de l'équation (3.2) et représente la vitesse de propagation du signal physique entre les hôtes références.

Chaque hôte référence L_i exécute un traceroute vers l'hôte cible τ pour découvrir la topologie du réseau (Fig. 5.3). Les différents routeurs sur le chemin ayant répondu par un message ICMP `TIME exceeded` sont pris en compte pour le calcul du chemin suivi par le traceroute. La position géographique de chaque routeur,

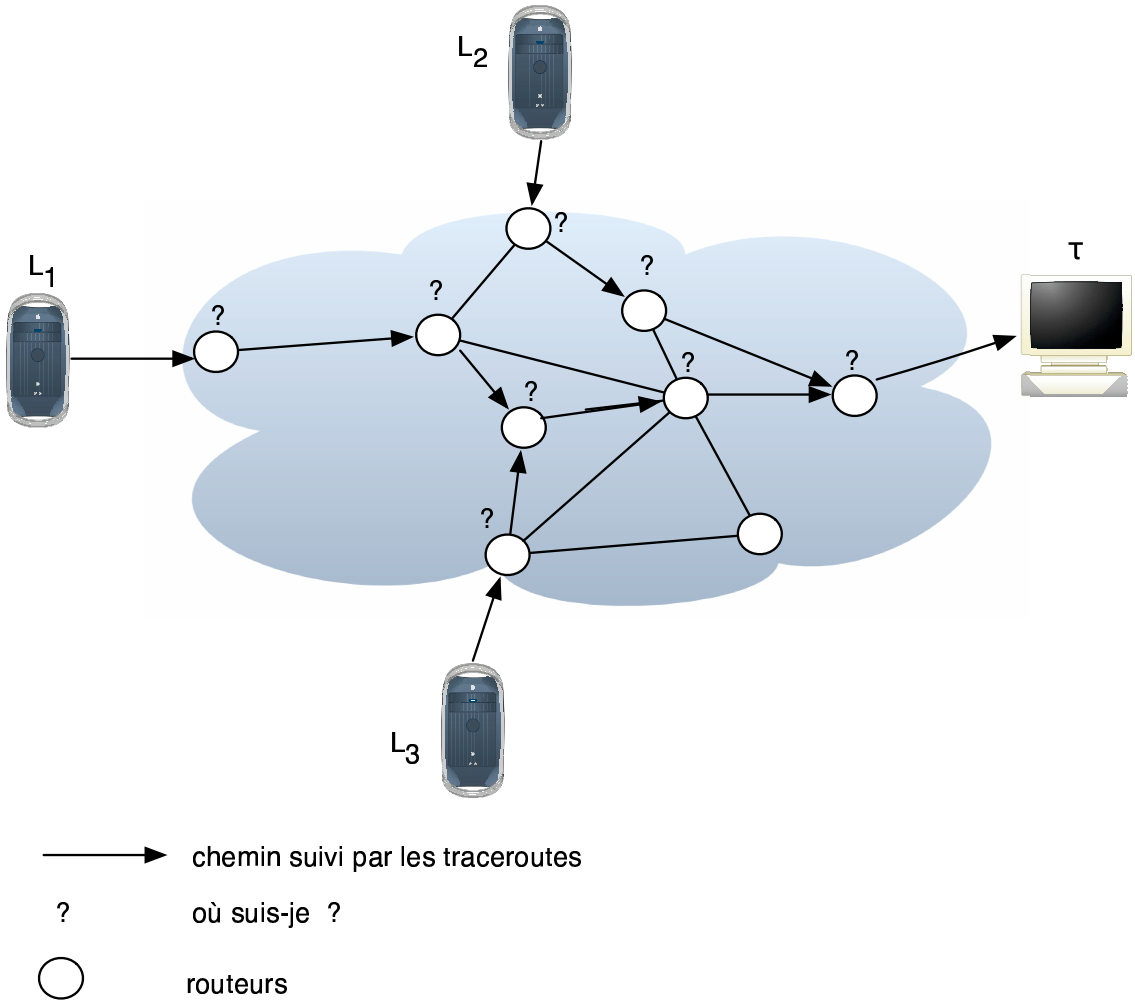


FIG. 5.3 – Découverte de topologie avec l'outil traceroute.

ayant répondu, est déterminée afin de calculer la distance parcourue par le traceroute. La somme des distances géographiques entre les routeurs adjacents sur le chemin donne l'estimation de la distance entre deux nœuds de bout en bout. Ainsi, la distance géographique suivie par le traceroute ente l'hôte référence L_i et l'hôte cible τ est donnée par

$$dist_{i\tau} = \sum_{k=0}^{n-1} dist(k, k+1), \tag{5.3}$$

où k représente le $k^{\text{ème}}$ routeur intermédiaire sur le chemin du traceroute (source, destination), n le nombre se sauts pour atteindre la destination τ, et *dist(k,k+1)* représente la distance géographique entre le nœud k et $k+1$.

Ainsi, à partir de l'équation (5.2), le délai de buffering b_k à chaque saut est obtenu grace à la formule

$$b_k = \Delta RTT_k - m_i \times dist(k-1, k)\cdot \tag{5.4}$$

Si on veut estimer b_k en utilisant l'équation 5.4, il est évident qu'il faut d'abord estimer la distance géographique entre chaque paire de routeurs adjacents sur le traceroute. Ainsi la connaissance de la position géographique de ces routeurs est nécessaire. Il est peu probable de connaître a priori la localisation géographique de tous les routeurs intermédiaires le long du traceroute (Fig. 5.3). Par conséquent, l'estimation du délai de buffering exige l'utilisation d'un service de géolocalisation pour identifier la position de l'ensemble des nœuds se trouvant sur le chemin qui mène vers la cible.

5.4 Évaluation de GeoBuD

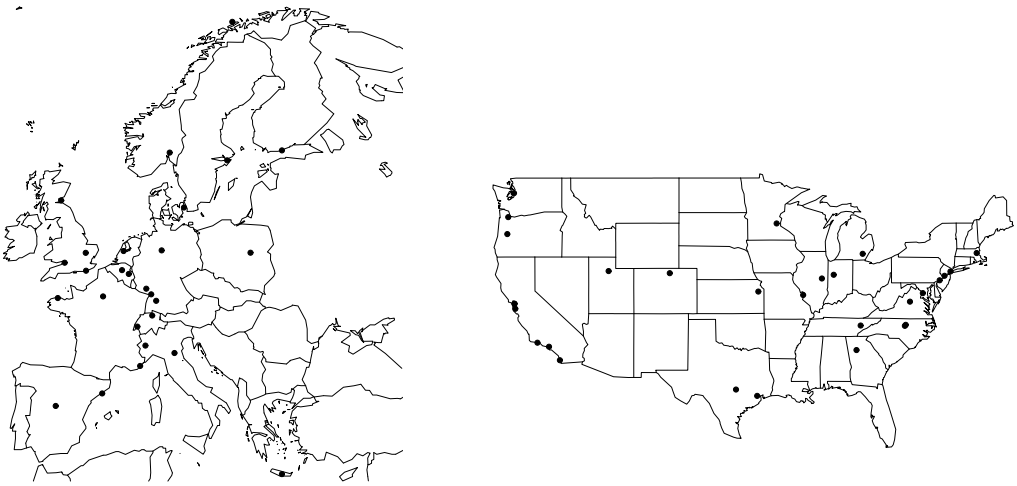
5.4.1 Paramètres expérimentaux

Pour estimer le délai de buffering b_k et la distance $dist(k, k+1)$ au niveau de chaque saut composant le chemin entre chaque paire (hôte référence, hôte cible), nous avons considéré deux ensembles de données :

- Premièrement, nous avons considéré 29 nœuds PlanetLab [92] localisés tous aux États Unis comme hôtes références. Ces hôtes références font des mesures de ping et de traceroute vers 87 nœuds AMP [89] qui représentent notre ensemble d'hôtes cibles situés aux États Unis. Ainsi, les données considérées dans notre premier échantillon sont composées par les mesures de traceroute et de ping faites par les hôtes références vers les hôtes cibles durant la journée du 17 Octobre 2005.
- Le deuxième ensemble de données est constitué par des nœuds tous situés en Europe. Il est ainsi formé par les mesures de traceroute et de de ping effectuées à partir de 27 nœuds de PlanetLab vers 57 nœuds RIPE [90] localisés en Europe. Les nœuds PlanetLab forment notre ensemble d'hôtes références et les nœuds RIPE celui des hôtes cibles. Ces mesures ont été effectuées durant la journée du 21 Novembre 2005.

La distribution géographique des hôtes références de nos deux ensembles de données est illustrée sur la figure (Fig. 5.4).

Dans l'approche CBG, pour inférer la position géographique d'un hôte cible, les hôtes références font des mesures de ping vers l'hôte cible. Les mesures de traceroute vers les mêmes hôtes cibles sont faites en parallèle avec les mesures de ping de CBG pour garder une équité entre les expériences par rapport à de probable variation de l'état du réseau.



(a) 27 hôtes références situés en Europe (b) 29 hôtes références situés aux U.S.

FIG. 5.4 – Distribution géographique des hôtes références (pas à la même échelle).

5.4.2 Estimation du délai de buffering

Les traceroutes exécutés par les hôtes références, situés aux États unis, vers les hôtes cibles AMP ont permis de découvrir un total de 1408 routeurs intermédiaires (sans les hôtes AMP). Nous avons pu localiser 1153 routeurs intermédiaires sur un nombre de 1408 routeurs découverts. 82% des routeurs ont été ainsi localisés.

Pour les hôtes localisés en Europe, sans les hôtes RIPE, nous avons localisé 1235 routeurs sur un total de 1328 routeurs découverts par nos hôtes références. Ainsi, 93% des routeurs ont été localisés.

Seuls les routeurs intermédiaires localisables sur le chemin, entre un hôte référence et l'hôte cible, sont pris en compte dans le calcul de l'estimation du chemin suivi par le traceroute. Il faut noter que la plupart des routeurs non localisés, sur le chemin du traceroute, sont situés à proximité de la source ou de la destination. Ainsi, l'erreur sur l'estimation du chemin suivi par le traceroute est donc probablement assez faible. Pour la localisation géographique de ces routeurs intermédiaires, nous avons utilisé le projet *GeoLIM* [91] qui est une implémentation de l'approche CBG. Nous avons également fait un recoupement des résultats obtenus par GeoLIM avec l'outil *rockettrace* qui fait parti du package *scriptroute* [107].

Certains sauts sur le traceroute peuvent subir l'effet de la congestion durant la période des mesures. Si cela survient, tout nœud ayant un RTT supérieur à celui du nœud qui le suit sur le chemin emprunté par le traceroute est écarté de l'estimation du délai de buffering. En effet, si nous tenons compte de ces nœuds, nous aurions à surestimer le délai de buffering mesuré au niveau de ces sauts.

Après avoir localisé les routeurs intermédiaires, nous calculons la valeur des b_k sur les différents segments du traceroute composés par ces routeurs en utilisant l'équation (5.2). Il faut noter que seuls les routeurs dont leur position géographique est connue sont pris en compte dans le calcul des b_k. Dans certains cas nous avons obtenu une estimation négative du délai de buffering b_k, ce qui présente manifestement un non-sens. Nous avons obtenu 21% de b_k négatifs représentant 4043 b_k sur un nombre total de 19172 b_k calculés en localisant les hôtes AMP. Pour les hôtes RIPE, le pourcentage de b_k négatifs obtenu est de 14% pour 11908 b_k trouvés. La plupart des b_k négatifs correspondent au cas où ΔRTT_{k+1} est très petit ou négatif. Ceci est dû à la variation de l'état du réseau le long du traceroute durant nos mesures. Ainsi, un b_k est considéré si et seulement si $\Delta RTT_{k+1} > 0$.

Pour chaque hôte référence L_i et la cible τ, le délai de buffering associé au chemin est

$$b_{i\tau} = \sum_{k=1}^{n-1} b_k, \tag{5.5}$$

où n représente le nombre de sauts intermédiaires le long du chemin sur le traceroute entre l'hôte référence L_i et l'hôte cible τ. Pour transformer les mesures de délai en distances géographiques surestimées, nous utilisons la formule ci-dessous qui dérive de l'équation (5.1) :

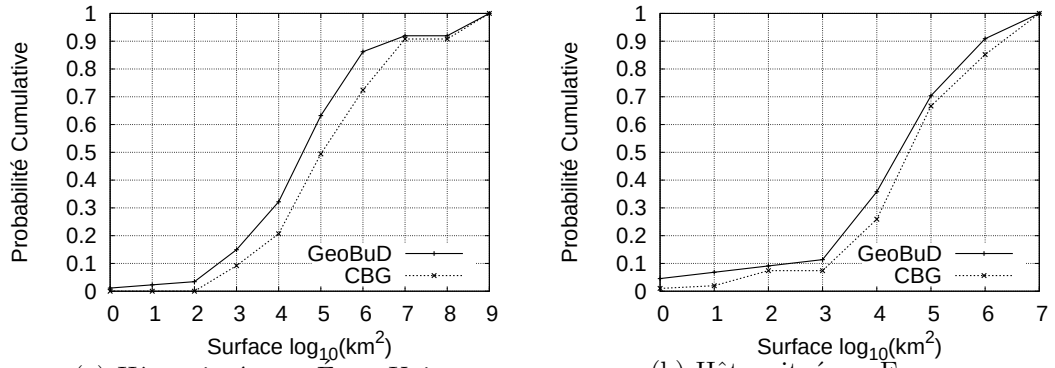
$$x_{i\tau} = \frac{y_{i\tau} - b_{i\tau}}{m_i}. \tag{5.6}$$

GeoBuD utilise la distance obtenue à partir de l'équation (5.6) pour localiser un hôte cible donné. Les distances géographiques obtenues par GeoBuD sont supposées être des bornes supérieures plus strictes sur la distance réelle que celles obtenues par la méthode CBG. En considérant ces nouvelles distances, et malgré le nombre de b_k négatifs, la zone de confiance obtenue avec la méthode GeoBuD est ainsi rétrécie. La précision de l'estimation de localisation est ainsi augmentée de même que la confiance qu'a le système en cette estimation (voir Section 5.4.3).

5.4.3 Réduction de la zone de confiance

La figure (Fig. 5.5) compare la probabilité cumulative de la taille de la zone de confiance obtenue par GeoBuD et CBG. L'axe des abscisses représente la surface de la zone de confiance associée à chaque estimation de localisation. L'axe des ordonnées montre la probabilité pour que l'estimation de la localisation ait une zone de confiance inférieure à une valeur x sur l'axe des abscisses.

En tenant compte du délai de buffering nous observons une nette amélioration



(a) Hôtes situés aux États Unis (b) Hôtes situés en Europe

FIG. 5.5 – Zone de confiance fournie par GeoBuD et CBG en km^2.

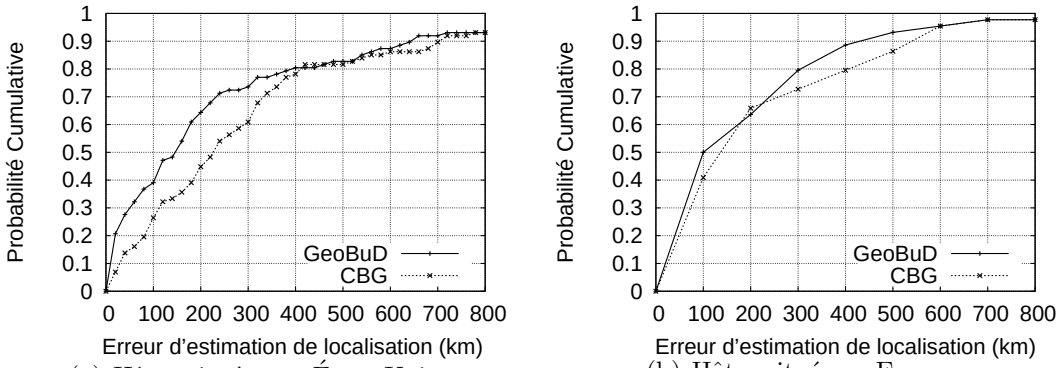
pour les zones de confiance inférieures à 10^7 km^2. En utilisant l'approche CBG, 72% des hôtes situés aux États Unis sont localisés avec une zone de confiance inférieure à 10^6 km^2. Pour cette même zone de confiance, nous localisons avec GeoBuD environ 86% des hôtes. En outre, avec CBG, 49% des hôtes cibles ont une zone de confiance inférieure à 10^5 km^2. Pour cette même zone de confiance, 63% des hôtes cibles sont localisés par GeoBuD. Pour les hôtes situés en Europe, GeoBuD en localise 10% avec une zone de confiance inférieure à 10^2 km^2. Notons qu'une surface de 10^5 km^2 équivaut à peu prés à la superficie d'un pays comme le Portugal ou bien d'un état des États Unis comme l'Indiana.

5.4.4 Erreur d'estimation de localisation

Nous nous attendons à ce que la réduction de la zone de confiance puisse avoir un impact sur l'erreur d'estimation de localisation de l'hôte cible. La figure 5.6 illustre la probabilité cumulative de l'erreur obtenue pour chaque estimation de localisation. L'erreur d'estimation est la différence entre la localisation géographique réelle de l'hôte cible et son estimation de localisation. Avec GeoBuD, 80% des hôtes situés aux États Unis présentent une erreur d'estimation moins importante comparée à CBG. Ainsi, l'erreur médiane est de 144 km en utilisant GeoBuD alors qu'elle est de 228 km pour CBG. Pour les hôtes situés en Europe, elle est de 100 km pour GeoBuD et de 137 km pour CBG.

5.4.5 Comparaison des distances géographiques estimées

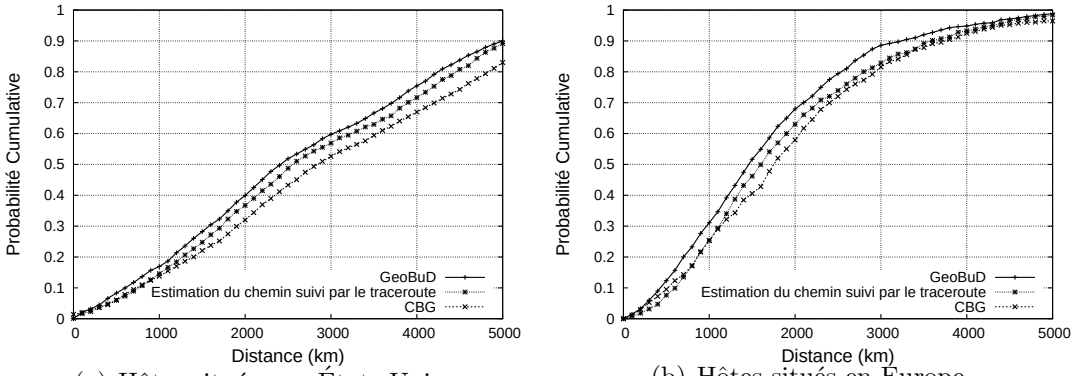
Pour améliorer la précision de l'estimation de localisation, il est nécessaire de réduire la zone de confiance dans laquelle l'hôte cible est supposé être. Le but est donc de surestimer de manière contrôlée la distance géographique afin d'obtenir



(a) Hôtes situés aux États Unis (b) Hôtes situés en Europe

FIG. 5.6 – Erreur d'estimation de localisation pour GeoBuD et CBG.

une zone de confiance la plus petite possible qui contient l'hôte cible. Il est donc nécessaire de vérifier que cette distance géographique surestimée obtenue est bien une borne supérieure de la distance réelle entre chaque hôte référence et l'hôte cible.



(a) Hôtes situés aux États Unis (b) Hôtes situés en Europe

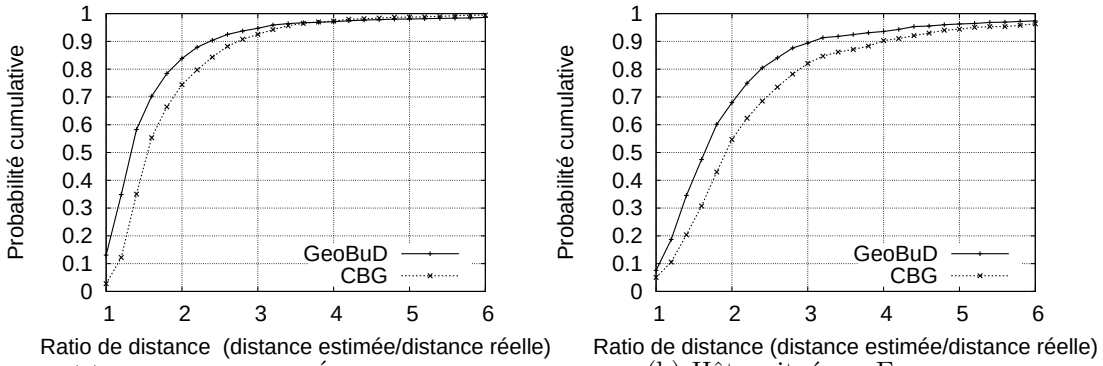
FIG. 5.7 – Comparaison des distances géographiques estimées.

La figure (Fig. 5.7) compare la probabilité cumulative de l'estimation de la distance géographique obtenue pour chaque paire (hôte référence, hôte cible) en utilisant GeoBuD, l'estimation du chemin suivi par le traceroute et CBG. Le chemin emprunté par le traceroute est obtenu en additionnant les distances géographiques entre les nœuds intermédiaires qui sont localisables le long du traceroute. Pour les distances supérieures à 1000 km, la figure (Fig. 5.7) montre nettement que les distances géographiques surestimées obtenues par CBG sont des bornes supérieures par rapport à la distance géographique réelle mais aussi

par rapport au chemin suivi par les traceroutes. Pour les distances inférieures à 1000 km, nous remarquons que la courbe de CBG est proche ou en dessous de celle de l'estimation du chemin suivi par le traceroute. Il faut noter que le chemin suivi par le traceroute est aussi une borne supérieure de la distance géographique réelle mesurée entre un hôte référence et l'hôte cible. Pour les distances de GeoBuD, nous observons sur la figure (Fig. 5.7) qu'elles sont des bornes supérieures plus strictes que celles de CBG ou à l'estimation du chemin suivi par le traceroute. Ainsi, GeoBuD fournit une plus petite zone de confiance que l'approche CBG.

Pour comprendre pourquoi GeoBuD est plus performant que CBG, il est nécessaire de rappeler comment CBG transforme les mesures de délai en distances géographiques surestimées. En effet, CBG se base sur un calibrage pour transformer les mesures de délai en distance géographique surestimée en définissant une valeur de b (voir équation (3.2)) qui dépend de la cible ayant la plus petite mesure de délai. Cette approche a l'avantage de fournir une borne supérieure conservatrice sur la distance, comme montré sur la figure (Fig. 5.7). Toutefois, l'inconvénient majeur comparé à GeoBuD est l'obtention de larges zones de confiance.

5.4.6 Comparaison entre la distance réelle et les distances géographiques estimées



(a) Hôtes situés aux États Unis (b) Hôtes situés en Europe

FIG. 5.8 – Comparaison du ratio entre les distances géographiques.

La figure (Fig. 5.8) montre la probabilité cumulative du ratio entre toutes les paires (hôte référence, hôte cible) de nos deux ensembles de données. Ce ratio représente le rapport entre l'estimation de distance, faite par CBG et GeoBuD, et la distance géographique réelle entre un hôte référence et un hôte cible. L'idée de la figure (Fig. 5.8) est de savoir comment chacune des deux approches CBG et GeoBuD surestiment la distance géographique réelle.

En effet, si nous savons que les distances géographiques sont surestimées avec un ratio supérieur à une constante nous pourrions procéder à un re-calibrage de nos estimations. Ce re-calibrage s'opère en divisant toutes les distances ainsi obtenues par cette constante. Malheureusement, nous remarquons que sur la figure 5.8, un petit pourcentage d'hôtes ont leur distance géographique estimée égale à leur distance géographique réelle (ratio = 1). Avec CBG, nous avons en effet 3 % des paires (hôte référence, hôte cible) dont la distance géographique réelle n'est pas surestimée. Quant à GeoBuD 13 % des paires sont dans ce cas (voir Fig. 5.8(a)). Sur la figure (Fig. 5.8(a)), nous avons 5% et 7% respectivement pour CBG et GeoBuD. Ainsi, si nous voulons faire un re-calibrage des distances estimées, nous serons amenés à faire une sous-estimation de la distance pour tous les hôtes ayant un ratio égal à 1. Cette sous-estimation conduirait à obtenir une zone de confiance vide en appliquant la multilatération avec ces distances re-calibrées [27]. La figure (Fig. 5.8) illustre aussi le fait que pour des ratios inférieurs à 4, GeoBuD surestime moins la distance que CBG.

5.5 Conclusion

La technique GeoBuD montre qu'en tenant compte du délai de buffering au niveau des sauts intermédiaires entre un hôte référence et un hôte cible, nous pouvons améliorer la précision de l'estimation de la localisation des hôtes dans l'Internet. En se basant sur des mesures faites avec l'outil traceroute, nous estimons le délai de buffering au niveau des routeurs intermédiaires sur le chemin entre nos hôtes références et la cible. Pour cela la position géographique des routeurs intermédiaires sur le chemin est nécessaire. La variation de l'état du réseau le long du traceroute fait que l'estimation du délai de buffering n'est pas toujours possible au niveau de chaque saut sur le chemin. Cela est matérialisé par l'existence de b_k négatifs lors de nos expérimentations. b_k représente le délai de buffering évalué au saut k sur le chemin. En associant cette estimation du délai de buffering avec une approche basée sur la multilatération (CBG), nous sommes capables de réduire la zone de confiance où l'hôte cible est localisé. Avec l'approche GeoBuD, 86% des hôtes situés aux États Unis sont localisés avec une zone de confiance inférieure à 10^6 km^2. Pour cette même zone de confiance nous localisons avec CBG 72% des hôtes. En outre, avec GeoBuD, 63% des hôtes cibles sont localisés avec une zone de confiance inférieure à 10^5 km^2. Quant aux hôtes situés en Europe, GeoBuD en localise 10% avec une zone de confiance inférieure à 10^2 km^2.

La réduction de la zone de confiance engendre une meilleure précision au niveau de l'estimation de localisation des hôtes cibles. Ainsi, avec GeoBuD, pour les hôtes cibles situés aux États Unis, l'erreur médiane est de 144 km tandis que

pour CBG elle est de 228 km. Pour les hôtes situés en Europe, elle est de 100 km pour GeoBuD et de 137 km pour CBG. Toutefois, ce gain de précision obtenu avec GeoBuD induit une charge supplémentaire au niveau du réseau avec la combinaison des mesures de ping et de traceroute faites par les hôtes références vers la cible. Pour pouvoir implémenter GeoBuD en temps réel, il faudrait avoir une cartographie du réseau Internet (position géographique des routeurs par exemple) sinon le temps de réponse nécessaire pour inférer la position géographique d'un hôte cible serait assez important.

Afin de réduire le nombre d'hôtes références considérés pour inférer la position d'un hôte cible, par conséquent le nombre de mesures engendrées, et le temps de réponse pour estimer la position d'un hôte cible, nous proposons une technique de géolocalisation hybride. Cette technique hybride repose sur les techniques de localisation basées sur des mesures passives et actives.

Chapitre 6

Vers un compromis entre mesures actives et mesures passives pour la localisation

Nous avons vu que les techniques de géolocalisation basées sur des mesures de délai engendraient beaucoup de trafic dans le réseau lors de l' inférence la localisation d'un hôte cible. En outre, le temps de réponse nécessaire à ces techniques pour inférer la localisation géographique de l'hôte cible est assez important. Il faut noter que, pour certaines applications dans l'Internet qui peuvent avoir besoin d'un service de géolocalisation, ce temps de réponse doit être équivalent au temps de chargement d'une page WEB recevant une requête, soit environ une à trois secondes en moyenne. Les techniques de géolocalisation basées sur des mesures de délai, ne peuvent pas fournir une réponse avec précision dans cet intervalle de temps.

Les applications commerciales de géolocalisation qui utilisent des bases de données peuvent être qualifiées de techniques passives [50, 51, 7, 10, 11, 6]. Ces bases de données contiennent des préfixes d'adresses IP et une information de localisation qui leur est associée. Bien qu'elles fournissent un temps de réponse assez rapide lorsqu'elles reçoivent une requête de localisation, la méthodologie et la précision de ces techniques de géolocalisation restent inconnues. Dans ce chapitre, nous proposons une technique hybride de géolocalisation qui utilise une base de données et des mesures de délai (la technique CBG) pour inférer la position des hôtes cibles dans l'Internet.

6.1 Système de géolocalisation hybride

6.1.1 Architecture hybride

La figure (Fig. 6.1) illustre les différents composants de notre système de géolocalisation hybride. Ce système peut être décomposé comme suit :

- Une base de données qui contient des préfixes d'adresses IP et leur propre information de localisation. Ces informations de localisation, selon les préfixes d'adresses IP répertoriés dans la base, peuvent être à l'échelle d'un pays, d'une région, d'une ville, ou bien sous forme de coordonnées géographiques (latitude, longitude).
- Un serveur où est implémenté l'heuristique qui détermine si des mesures doivent être faites vers l'hôte cible, et si tel est le cas, avec quel ensemble d'hôtes références (landmarks).
- Des mesures actives (mesures de délai) qui sont faites à partir d'un ensemble d'hôtes références prédéfini.

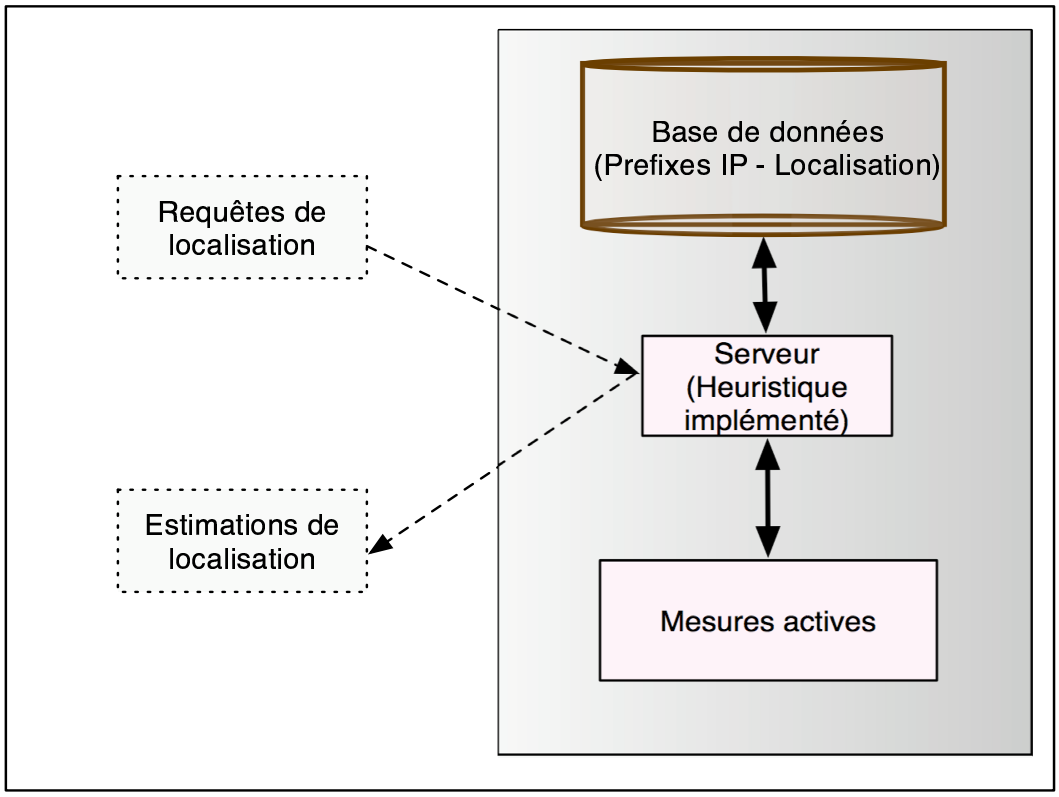


FIG. 6.1 – Architecture d'un système de géolocalisation hybride.

TAB. 6.1 – Bases de données des préfixes d'adresses IP.

table 1	
préfixe d'adresses IP	**loc id**

table 2						
loc id	**pays**	**région**	**ville**	**code postal**	**latitude**	**longitude**

Le processus de localisation d'un hôte cible à partir de notre système hybride de géolocalisation est expliqué plus en détail dans la section 6.2.

6.1.2 Structure de la base de données utilisée

Dans un premier temps, quand une requête de localisation arrive au niveau du système illustré sur la figure Fig. 6.1, le serveur interroge la base de données pour connaître la position géographique de cet hôte cible. La base de données étudiée ici, est celle qu'utilise *GeoIP* [10]. Nous avons utilisé la version commerciale que fournit *MaxMind* [108]. La base de données de GeoIP est la plus utilisée actuellement.

En effet, la base de données contient des préfixes d'adresses IP et leur information de localisation. Elle est constituée de deux tables, *table 1* et *table 2*, comme illustré dans le tableau (Tab. 6.1). Dans la base de données, chaque entrée de la table 1 contient la valeur du préfixe d'adresses IP et son numéro d'identification (voir tableau (Tab. 6.1)). Au niveau de la table 2, chaque entrée contient l'identifiant du préfixe d'adresses IP et les informations de localisation qui lui sont associées suivant une granularité plus fine (voir tableau (Tab. 6.1)).

Ensuite, une tabulation exhaustive, comme dans [50, 51, 10, 6], permet de trouver, s'il existe dans la base de données, le préfixe d'adresses IP auquel appartient l'hôte cible. En connaissant le préfixe d'adresses IP auquel appartient l'hôte cible, nous déduisons, à partir de la table 2 de la base de données, la localisation géographique de l'hôte cible. Il faut noter que, la tabulation exhaustive donne comme réponse le préfixe d'adresses auquel appartient la cible ayant le plus long préfixe dans la base de données.

Ainsi, après avoir obtenu la localisation du préfixe d'adresses IP de l'hôte cible, nous appliquons l'heuristique qui permet de définir un sous-ensemble, dans notre ensemble d'hôtes références $\mathcal{L}$, qui va procéder à la localisation de l'hôte cible. Par contre, si l'adresse IP de l'hôte cible n'appartient à aucun préfixe d'adresses IP de la base de données, *i.e* le préfixe d'adresses IP n'est pas enregistré dans la base, alors la localisation de l'hôte cible se fait avec tous les hôtes références de notre ensemble.

6.2 Heuristique du choix des hôtes références

Comme illustré au niveau de l'architecture de notre système hybride de géolocalisation (Fig. 6.1), le serveur implémente une heuristique qui lui permet de choisir l'ensemble des hôtes références qui feront les mesures vers l'hôte cible. Pour ce faire, le serveur envoie une requête à la base de données pour connaître la localisation géographique de cet hôte cible. Si cette localisation existe, alors la base de données lui renvoie la latitude et la longitude du préfixe d'adresses IP auquel appartient cet hôte cible.

L'heuristique consiste, pour un nombre d'hôtes références fixé, à choisir les hôtes références les plus proches à l'hôte cible en terme de distance géographique. Il faut noter que la position géographique (latitude, longitude) de tous les hôtes références qui constituent notre ensemble $\mathcal{L}$ est connue. Connaissant la position géographique du préfixe d'adresses de l'hôte cible, grace à la base de données, et la position géographique des hôtes références, nous calculons la distance géographique entre les hôtes références et l'hôte cible. En se basant sur [109], la distance géographique, entre chaque hôte référence L_i et l'hôte cible τ, est donnée par

$$\beta = \sqrt{\left(\sin\left(\frac{lat_i - lat_\tau}{2}\right)\right)^2 + \cos(lat_i) \times \cos(lat_\tau) \times \left(\sin\left(\frac{lon_\tau - lon_i}{2}\right)\right)^2} \tag{6.1}$$

$$\hat{dist}_{i\tau} = 6371 \times 2 \times \arcsin(\beta) \tag{6.2}$$

Dans l'équation 6.1, lat_i, et lon_i représentent la latitude et la longitude, exprimées en radian, de l'hôte référence L_i ; lat_τ et lon_τ représentent la latitude et la longitude, exprimées en radian, de l'hôte cible τ. La distance géographique, exprimée en km, entre l'hôte référence L_i et l'hôte cible τ, est obtenue à partir de l'équation 6.2. Le terme 6371, présent dans l'équation 6.2, représente le rayon de la terre [110]. En effet, l'expression $2 \times \arcsin(\beta)$ fournit la distance géographique en radian. Dans la section 6.3, ce sont les distances géographiques exprimées en km qui sont considérées.

Pour l'hôte cible τ, nous obtenons le vecteur de distance

$$D_\tau = [\hat{dist}_{1\tau}, \hat{dist}_{2\tau}, \ldots, \hat{dist}_{K\tau}], \tag{6.3}$$

où K représente le nombre d'hôtes références total de notre ensemble $\mathcal{L}$ et $\hat{dist}_{i\tau}$ représente la distance géographique, en km, calculée entre l'hôte référence L_i et la cible τ pour $1 \leq i \leq K$.

Si nous fixons par exemple un nombre n d'hôtes références parmi les K que compte notre ensemble $\mathcal{L}$, les n hôtes références ayant les plus petits $\hat{dist}_{i\tau}$, $1 \leq i \leq n$, sont choisis pour inférer la position de l'hôte cible. Nous montrons dans la section 6.3.2, que pour un nombre réduit d'hôtes références, avec cette heuristique, nous obtenons de meilleurs résultats, de surcroît avec moins de mesures injectées dans le réseau.

6.3 Évaluation

6.3.1 Paramètres expérimentaux

Pour évaluer notre heuristique, nous avons considéré un ensemble de données constitué par des hôtes AMP [89] et des hôtes RIPE [90] comme hôte cible. Ces hôtes sont au nombre de 127 et sont localisés essentiellement aux États Unis et en Europe. La principale raison de cette restriction est due au fait que nous avons besoin d'hôtes cibles dont on connaît la localisation géographique pour pouvoir évaluer l'erreur d'estimation. Ainsi, seuls les hôtes AMP et RIPE fournissent cette exigence.

Nous avons utilisé la base de données qu'emploie *GeoIP* [10]. GeoIP est une technologie propriétaire, un outil commercial, appartenant à l'organisation *MaxMind* [108]. Nous avons utilisé la version commerciale de GeoIP. La différence entre la version gratuite et la version commerciale de GeoIP est au niveau de la granularité de l'information de localisation fournie. Dans la base gratuite de GeoIP, l'information de localisation est fournie seulement à l'échelle d'un pays. Cette base de données est formée par des préfixes d'adresses IP et chaque préfixe d'adresses IP possède une information de localisation suivant plusieurs granularités comme le pays, la région, la ville ou latitude/longitude (Tab. 6.1). La base de données contient 1 873 596 préfixes d'adresses IP et le masque de ces préfixes d'adresse varie entre 8 et 32.

Pour inférer la localisation de ces hôtes cibles, nous avons considéré un ensemble de 74 nœuds PlanetLab [92] comme hôtes références. Les points illustrés sur la figure (Fig. 6.2) représentent la distribution géographique de ces hôtes références.

Nous avons fait les mesures de délai entre les hôtes références et les hôtes cibles durant la semaine du 10 au 15 juillet 2006. Chaque hôte référence exécuté des pings vers les hôtes cibles pour mesurer le délai entre eux. Chaque mesure de ping, entre un hôte référence L_i et un hôte cible τ, est composé de 10 paquets sondes envoyés par intervalle de 1 seconde. Nous espaçons les paquets sondes pour que nos mesures ne soient pas considérées comme des attaques de déni de service. Chaque paquet envoyé a une taille de 1024 Ko. Seul le RTT minimum est considéré pour chaque mesure de ping. Il est plus vraisemblable, que le RTT

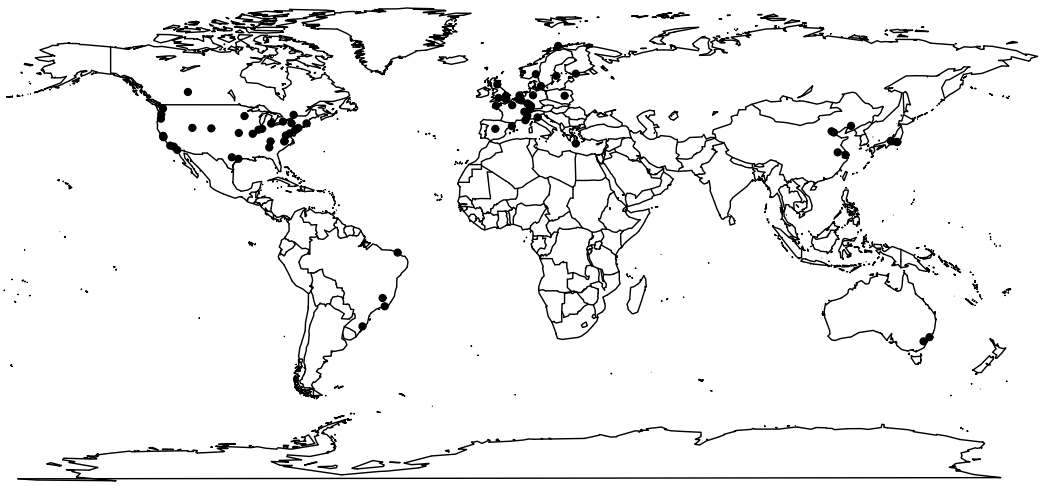


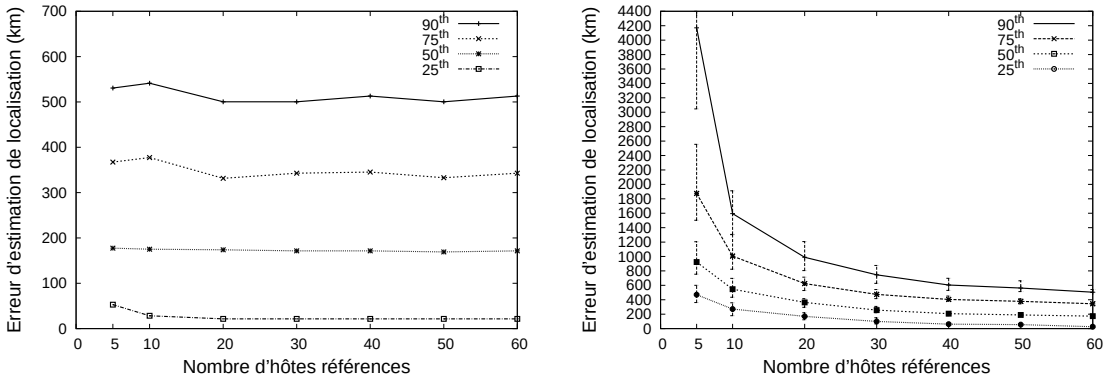
FIG. 6.2 – Distribution géographique des hôtes références.

minimum reflète le mieux le délai de propagation et qu'il soit le moins assujetti aux congestions et autres sources de distorsions. Ainsi, nous avons considéré le RTT minimum entre chaque hôte référence et chaque hôte cible de notre ensemble de données. Pour inférer la localisation géographique des hôtes cibles nous utilisons l'approche CBG décrite dans le chapitre 3.

6.3.2 Résultats

La figure (Fig. 6.3) montre l'impact du choix des hôtes références et de leur nombre sur les performances de CBG. Le choix des hôtes références, pour localiser les hôtes cibles, se fait soit de manière aléatoire ou, suivant l'heuristique étudiée dans la section 6.2.

La figure (Fig. 6.3(a)) montre différents centiles de l'erreur d'estimation de localisation obtenue en fonction du nombre d'hôtes références considérés en utilisant la technique CBG. Le choix des hôtes références, pour un nombre k d'hôtes références fixé, se fait suivant l'heuristique présentée dans la section 6.2. L'axe des abscisses représente le nombre d'hôtes références choisi dans notre ensemble composé de 74 hôtes références, pour inférer la localisation d'un hôte cible. Le nombre d'hôtes références varie entre 5 et 60. L'axe des ordonnées représente l'erreur d'estimation de localisation obtenue pour un nombre k d'hôtes références utilisés. Par exemple, la courbe qui montre les 90ème de centiles, représente l'erreur d'estimation de localisation, où la courbe de la fonction de probabilité cumulative de l'erreur moyenne d'estimation de localisation rencontre le point dont la pro-



(a) Choix des hôtes références par heuristique (b) Choix aléatoire des hôtes références

FIG. 6.3 – Erreur d'estimation de localisation en fonction du nombre d'hôtes références.

babilité est 90%. Nous remarquons qu'à partir de 20 hôtes références considérés (Fig. 6.3(a)), l'erreur d'estimation de localisation reste stable. Toutefois, pour la courbe des 90 et 75 centiles nous notons une légère augmentation de l'erreur d'estimation de localisation si l'on augmente le nombre d'hôtes références considérés. Ceci est certainement dû à la présence de "bruit" (distorsions) au niveau de nos mesures de délai induit par les hôtes références ajoutés, et qui sont un peu éloignés de la cible. Plus les hôtes références sont proches géographiquement de la cible meilleur est l'estimation de localisation (Fig. 6.3(a)). En ne considérant que les 20 hôtes références les plus proches des hôtes cibles, 50% des hôtes cibles sont localisés avec une erreur inférieure à 175 km (voir Fig. 6.3(a)).

La figure (Fig. 6.3(b)) montre différents centiles de l'erreur d'estimation de localisation en fonction du nombre d'hôtes références considérés. Le choix des hôtes références, pour un nombre k d'hôtes références fixé, s'est fait de manière aléatoire. Nous avons considéré 30 échantillons de mesures pour chaque k hôtes références choisis. Le nombre d'hôtes références considéré varie entre 5 et 60. Toutefois, comme le nombre de possibilité de placement des hôtes références devient de plus en plus important lorsque k augmente, nous n'avons pas considéré toutes les façons de choisir k hôtes références dans chaque ensemble de données.

Les barres sur la figure (Fig. 6.3(b)) représentent les intervalles de confiance, pour l'ensemble des échantillons de mesures considérées, pour un nombre d'hôtes références fixé. Nous remarquons qu'à partir de 30 hôtes références, l'erreur d'estimation de localisation est pratiquement stable. Toutefois, avec un choix aléatoire des hôtes références, l'erreur d'estimation de localisation reste assez importante Ainsi, en considérant 30 hôtes références choisis aléatoirement, 50% des hôtes cibles sont localisés avec une erreur d'estimation inférieure à 400 km.

En considérant notre heuristique, pour ce même nombre d'hôtes cibles et d'hôtes références, l'erreur d'estimation de localisation est inférieure à 175 km (Fig. 6.3(a)).

Ainsi, en considérant un nombre restreint d'hôtes références pour localiser les hôtes cibles, nous réduisons le temps de traitement nécessaire à CBG pour traiter une requête de géolocalisation. Par conséquent le temps de réponse est fortement diminué.

6.4 Conclusion

Dans ce chapitre, nous avons proposé et évalué une technique utilisant à la fois des mesures actives et passives pour inférer la position géographique des hôtes Internet. Nous avons mis en place un système hybride de géolocalisation qui permet d'utiliser une base de données (mesure passive) que l'on associe à des mesures de délai (mesure active). Ainsi, grace à la base de données, il est possible d'inférer la position de l'hôte cible avec les hôtes références qui lui sont le plus proche géographiquement. En effet, l'heuristique que nous avons développé permet de choisir les hôtes références les plus proches géographiquement de la cible.

Les résultats obtenus montrent que, si l'on choisi les hôtes références les plus proches géographiquement de la cible, pour le localiser, nous obtenons une meilleure estimation de localisation comparé au choix aléatoire des hôtes références. En ne considérant que les 20 hôtes références les plus proches des hôtes cibles (choix par heuristique), 50% des hôtes cibles sont localisés avec une erreur inférieure à 175 km. En outre, 20 hôtes références suffisent pour stabiliser l'erreur d'estimation de localisation. La combinaison de la base de données avec les mesures de délai permet de réduire le nombre d'hôtes références utilisés et par ailleurs réduire le trafic injecté dans le réseau ainsi que le temps de réponse pour inférer la position de la cible.

Cependant, la mise-à-jour des bases de données reste difficile à faire. Malgré l'utilisation de la base de données pour choisir les hôtes références les plus proches à l'hôte cible, nous notons une erreur d'estimation au niveau de la localisation des hôtes cibles. Nous envisageons de déployer cette technique hybride sur PlanetLab mais aussi d'évaluer la précision de la base de données.

Chapitre 7

Conclusion

Avec la prolifération des moteurs de recherche et de distributeurs de contenu dans l'Internet, la géolocalisation et ses services associés sont voués actuellement à un essor économique important et pour encore bien longtemps. Les techniques de géolocalisation permettent ainsi d'inférer la position des hôtes dans l'Internet uniquement à partir de leur adresse IP. Dans cette thèse, nous nous sommes focalisés sur les techniques de géolocalisation basées sur des mesures de délai pour inférer la position des hôtes cibles dans l'Internet.

Les techniques de localisation géographique basées sur des mesures de délai exploitent une possible corrélation entre délai et distance géographique. La technique CBG, proposée dans le chapitre 3, incorpore un algorithme d'auto-calibration entre les hôtes références (hôte dont on connaît la position géographique) pour capturer la relation existant entre délai et distance géographique. En effet, cette auto-calibration, tente d'enlever les différentes distorsions pouvant réduire la corrélation entre délai et distance géographique, et elle se fait entre les hôtes références. L'auto-calibration, grace à l'utilisation de la bestline, permet de capturer la meilleure relation pouvant exister entre délai et distance géographique dans le réseau. En appliquant la multilatération, CBG fournit un espace continu d'endroits, où on peut localiser les hôtes cibles, contrairement aux précédentes techniques de géolocalisation basées sur des mesures de délai. Avant d'appliquer la multilatération dans l'Internet, CBG transforme les mesures de délai, entre les hôtes références et l'hôte cible, en distances géographiques estimées par le biais des bestlines des hôtes références. Ces distances géographiques étant toujours "surestimées" permettent à la technique CBG d'obtenir une zone d'intersection dans laquelle la position de l'hôte cible est inférée. CBG est capable d'associer aussi à chaque estimation de localisation de l'hôte cible une zone de confiance. Cela

permet aux applications, qui utilisent CBG, d'évaluer la fiabilité de l'estimation par rapport à leurs exigences.

Nous avons montré que le manque de précision de la technique CBG était dû à la présence de sources de distorsions dans le réseau. Ces sources de distorsions ajoutent un temps supplémentaire dans les mesures de délai et sont dues à la présence de goulots d'étranglement, de "localized delay", de chemins partagés, et à la non linéarité du chemin entre les hôtes Internet ("path inflation"). Ainsi, nous avons estimé l'impact du délai de buffering sur les mesures de délai entre les hôtes références et les hôtes cibles dans le chapitre 5. Pour ce faire, nous avons utilisé l'outil traceroute. Les résultats obtenus montrent qu'en supprimant ce délai de buffering, avant de transformer les mesures de délai en distance géographiques estimées, nous réduisons la taille des zones de confiance associées à chaque estimation de localisation. En outre, la précision de l'estimation de localisation se trouve améliorée. Toutefois, cette proposition génère beaucoup de mesures dans le réseau et elle est assez complexe à déployer.

Pour réduire le nombre d'hôtes références considérés pour inférer la localisation des hôtes cibles, une technique hybride, associant l'utilisation d'une base de données et des mesures de délai, a été proposée dans le chapitre 6. La base de données utilisée est formée par des préfixes d'adresses IP et leur information de localisation. Une heuristique, qui permet de choisir les hôtes références les plus proches géographiquement de l'hôte cible, a été développée. Les résultats obtenus montrent que 20 hôtes références suffisent pour obtenir une bonne estimation de localisation. En outre, nous réduisons le nombre de mesures générées par les hôtes références pour localiser un hôte cible.

7.1 Perspectives

Actuellement, le développement fulgurant des techniques de géolocalisation peut mener à des questionnements divers. Que font les gens par rapport aux informations de localisation obtenues à partir des internautes. Un problème d'éthique se pose concernant une possible violation de la vie privée des internautes. Ainsi, le groupe de travail geopriv ("*Geographic Location/Privacy*") [111] à l'IETF ("*Internet Engineering Task Force*") est chargé de mettre en place des politiques de sécurité concernant les informations de localisation des utilisateurs. Une standardisation des informations de localisation fournies est aussi envisagée.

Nous explorons des alternatives pour surmonter la limitation due à la présence de proxies et dont souffre toutes les techniques de géolocalisation dans l'Internet qui se basent uniquement sur les adresses IP. Ces techniques de géolocalisation sont celles basées sur les bases de données Whois, le traceroute, et les mesures de délai. Nous envisageons de tenir compte des différents sources de distorsions

élucidées dans cette thèse, et pouvant occasionner des imprécisions au niveau des mesures de délai et par conséquent une mauvaise estimation de localisation des hôtes cibles.

La corrélation entre délai et distance géographique peut "tomber" si le dernier lien ("*last-mile*") vers l'hôte cible est un lien à bas débit (par exemple modem, ou lien satellite) occasionnant des délais supplémentaires. Ainsi, nous envisageons de prendre compte du type de lien utilisé par les hôtes cibles. Par exemple, si nous savons que l'hôte cible utilise un modem, qui peut être détecté par une estimation de la bande passante ou par un traceroute, nous pouvons demander à la technique de localisation, soit d'estimer la localisation du dernier routeur sur le chemin vers la cible, soit de tenir compte de cette limitation.

Un déploiement de la technique hybride de géolocalisation proposée dans le chapitre 6 est envisagé. Dans ce déploiement, en plus du choix des hôtes références suivant leur proximité géographique par rapport à la cible, nous envisageons d'ajouter une nouvelle heuristique qui permet de vérifier si la localisation fournie par la base de données est précise. Dans ce cas, aucune mesure ne sera faite. Par contre, si l'information de localisation fournie est imprécise, on fera des mesures actives, et ensuite, on vérifiera sa fiabilité pour faire une mise-à-jour de la base de données.

Publications

- **Journaux internationaux**
  - Bamba Gueye, Artur Ziviani, Mark Crovella, Serge Fdida
    "Constraint-Based Geolocation of Internet Hosts", *IEEE/ACM Transaction on Networking, IEEE/ACM Press, ISSN :1063-6692, à paraître Décembre 2006.*

- **Conférences Internationales avec comité de lecture**
  - Bamba Gueye, Steve Uhlig, Artur Ziviani, Serge Fdida
    "Leveraging Buffering Delay Estimation for Geolocation of Internet Hosts", *in Proc. IFIP Networking, Coimbra, Portugal, Mai 2006, Lectures Notes in Computer Science (LNCS) 3976, pp. 319-330.*

  - Bamba Gueye, Artur Ziviani, Mark Crovella, Serge Fdida
    "Constraint-Based Geolocation of Internet Hosts", *in Proc. ACM/SIGCOMM Internet Measurement Conference (IMC), Taormina, Sicily, Italy, Octobre 2004, pp. 288–293.*

  - Bamba Gueye, Artur Ziviani, Serge Fdida, José F.D. Rezende, Otto Carlos M.B. Duarte
    "Two-Tier Geographic Location of Internet Hosts", *in Proc. IEEE International Conference on High Speed Networks and Multimedia Communications (HSNMC), Toulouse, France, Juin 2004, Lectures Notes in Computer Science (LNCS) 3079, pp. 730-739.*

- **Conférences Francophones avec comité de lecture**
  1. Bamba Gueye, Artur Ziviani, Mark Crovella, Serge Fdida
     "Techniques de localisation géographique d'hôtes dans l'Internet", *in Proc. Septièmes Rencontres Francophones sur les aspects Algorithmiques des Télécommunications (ALGOTEL), Presqu'Ile de Giens, France, Mai 2005.*
  2. Bamba Gueye, Artur Ziviani, Mark Crovella, Serge Fdida
     "Vers la Localisation Géographique d'Hôtes dans l'Internet basée sur la Multilatération", *in Proc. Colloque Francophone sur L'Ingénierie des Protocoles (CFIP), Bordeaux, France, Mars 2005, pp 415–431.*

Bibliographie

[1] comScore Networks Inc, http://www.comescore.com/.

[2] Google Inc, http://www.google.com/.

[3] Yahoo ! Inc, http://www.yahoo.com/.

[4] Microsoft Corporation, http://www.microsoft.com/.

[5] Forrester Research Inc, http://www.forrester.com/.

[6] *GeoURL*, http://www.geourl.org/.

[7] *GeoNetMap*, Geobytes, Inc., http://www.geobytes.com/GeoNetMap.htm.

[8] *WhereIsIP*, Qwerks, Inc., http://www.jufsoft.com/whereisip/.

[9] *ActiveTarget*, RegSoft.com Inc., http://www.activetarget.com/.

[10] *GeoIP*, MaxMind LLC, http://www.maxmind.com/geoip/.

[11] *GeoPoint*, Quova Inc., http://www.quova.com/.

[12] *Net World Map*, http://www.networldmap.com/.

[13] *IP Address to Latitude/Longitude*, University of Illinois at Urbana-Champaign, http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll/.

[14] *OASIS*, CoralCDN, http://www.coralcdn.org/oasis/.

[15] *Le forum des droits sur l'Internet*, 2005, http://www.internet.gouv.fr/article.php3?id_article=1855.

[16] P. Enge and P. Misra, "Special issue on global positioning system," *Proc IEEE*, vol. 87, no. 1, pp. 3–15, Jan. 1999.

[17] E. Ermel, "Localisation et routage géographique dans les réseaux sans fil hétérogènes," Ph.D. dissertation, Université Pierre et Marie Curie (Paris 6), Paris, France, June 2004.

[18] R. Jain, A. Puri, and R. Sengupta, "Geographical routing using partial information for wireless ad hoc networks," *IEEE Personal Communications*, vol. 8, no. 1, pp. 48–57, Feb. 2001.

[19] T. Camp, J. Boleng, B. Williams, L. Wilcox, and W. Navidi, "Performance comparison of two location based routing protocols for ad hoc networks," in *Proc. IEEE INFOCOM*, New York, NY, USA, June 2002.

[20] M. J. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proc. ACM/SIGCOMM Internet Measurement Conference*, Berkeley, CA, USA, Oct. 2005.

[21] A. Lakhina, J. W. Byers, M. Crovella, and I. Matta, "On the geographic location of Internet resources," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 934–948, Aug. 2003.

[22] S. H. Yook, H. Jeong, and A. Barabási, "Modeling the Internet's large-scale topology," *Proc. of the National Academy of Sciences (PNAS)*, vol. 99, pp. 13 382–13 386, Oct. 2002.

[23] S. Banerjee, T. G. Griffin, and M. Pias, "The interdomain connectivity of PlanetLab nodes," in *Proc. the Passive and Active Measurement Workshop – PAM*, ser. Lecture Notes in Computer Science (LNCS) 3015, Antibes Juan-les-Pins, France, Apr. 2004.

[24] L. Subramanian, V. Padmanabhan, and R. Katz, "Geographic properties of Internet routing," in *Proc. USENIX*, Monterey, CA, USA, June 2002.

[25] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Vers la localisation géographique d'hôtes dans l'internet basée sur la multilatération," in *CFIP 05*, Bordeaux, France, Mar. 2005.

[26] ——, "Constraint-based geolocation of Internet hosts," in *Proc. ACM/SIGCOMM Internet Measurement Conference – IMC*, Taormina, Sicily, Italy, Oct. 2004, pp. 288–293.

[27] ——, "Constraint-based geolocation of internet hosts," *IEEE/ACM Transactions on Networking*, 2006, to appear.

[28] B. Gueye, S. Uhlig, A. Ziviani, and S. Fdida, "Leveraging buffering delay estimation for geolocation of Internet host," in *Proc. IFIP Networking Conference*, ser. Lecture Notes in Computer Science (LNCS), Coimbra, Portugal, May 2006, pp. 319–330.

[29] *GALILEO – European Satellite Navigation System*, The European Union, http://europa.eu.int/comm/dgs/energy_transport/galileo/.

[30] *GLONASS – Russian Satellite Navigation System*, http://www.glonass-center.ru/.

[31] *Beidou– Russian Satellite Navigation System*, http://www.cast.ac.cn/en/printpage.asp?ArticleID=36.

[32] A. Harter and A. Hopper, "A distributed location system for the active office," *IEEE Network*, vol. 8, no. 1, pp. 62–70, Jan. 1994.

[33] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in *Proc. of ACM/IEEE International Conference on Mobile Computing and Networking – MobiCom*, Aug. 2000.

[34] P. Bahl and V. N. Padmanabhan, "RADAR : An in-building RF-based user location and tracking system," in *Proc. IEEE INFOCOM*, Tel-Aviv, Israel, Mar. 2000.

[35] M. Rahnema, "Overview of the GSM system and protocol architecture," *IEEE Communications Magazine*, vol. 31, no. 4, pp. 92–100, Apr. 1993.

[36] T. Clausen and P. Jacquet, *Optimized Link State Routing*, Oct. 2003.

[37] D. Johnson, D. Maltz, and J. Broch, *DSR The Dynamic Source Routing Protocol for Multihop Wireless Ad Hoc Networks.* Addison-Wesley, 2001, ch. 5, pp. 139–172.

[38] C. E. Perkins and E. M. Royer, "Ad hoc on demand distance vector (AODV) routing," in *Proc IEEE Workshop Mobile Computing Systems and Apllications(WMCSA)*, Feb. 1999.

[39] S. Giordano, I. Stojmenovic, and L. Blazevie, "Position based routing algorithms for ad hoc networks : a taxonomy," 2001.

[40] Y.-B. Ko and N. H. Vaidya, "Location-aided routing (LAR) in mobile ad hoc networks," in *Proc. ACM MobiCom'98*, Dallas, TX, USA, Oct. 1998.

[41] J. C. Navas and T. Imielinski, "GeoCast – geographic addressing and routing," in *Mobile Computing and Networking*, 1997, pp. 66–76.

[42] T. Imielinski and J. Navas, *Gps-based addressing and routing*, Mar. 1996.

[43] B. Karp and H. T. Kung, "GPSR : Greedy perimeter stateless routing for wireless networks," in *Proc. ACM MobiCom*, Boston, MA, USA, Aug. 2000.

[44] P. Bose, P. Morin, I. Stojmenovic, and J. Urrutia, "Routing with guaranteed delivery in ad hoc wireless networks," *Wireless Networks*, vol. 7, no. 6, pp. 609–616, 2001.

[45] K. R. Gabriel and R. R. Sokal, "A new statistical approach to geographic variation analysis," *Systematic Zoology*, vol. 18, no. 3, pp. 259–278, Sept. 1969.

[46] C. Davis, P. Vixie, T. Goodwin, and I. Dickinson, "A means for expressing location information in the domain name system," *Internet RFC 1876*, Jan. 1996.

[47] B. Huffaker, M. Fomenkov, D. J. Plummer, D. Moore, and k claffy, "Distance metrics in the Internet," in *Proc. IEEE International Telecommunications Symposium - ITS*, Natal, Brazil, Sept. 2002.

[48] D. Moore, R. Periakaruppan, J. Donohoe, and k.c. Claffy, "Where in the world is netgeo.caida.org ?" in *Proc. of INET*, Yokohama, Japan, July 2000.

[49] M. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proc. ACM/SIGCOMM Internet Measurement Conference*, Berkeley, CA, USA, Oct. 2005.

[50] Akamai Inc, http://www.akamai.com/.

[51] Digital Island Inc, http://www.digitalisland.com/.

[52] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *SIGCOMM*, San Diego, CA, USA, Aug. 2001.

[53] *VisualRoute*, Visualware Inc., http://www.visualware.com/visualroute/.

[54] *GTrace*, CAIDA, http://www.caida.org/tools/visualization/gtrace/.

[55] Sarangworld Traceroute Project, http://www.sarangworld.com/TRACEROUTE/.

[56] V. Jacobson, *Traceroure Software*, 1999, ftp://ftp.ee.lbl.gov/traceroute.tar.Z.

[57] B. Krishnamurthy and J. Wang, "On network-aware clustering of web clients," in *SIGCOMM*, 2000, pp. 97–110.

[58] K. Lougheed and Y. Rekhter, *A Boder Gatewey Protocol*, June 1990.

[59] B. Halabi, *Internet Routing Architectures*. Cisco Press, 1997.

[60] J. T. Moy, *OSPF : Anatomy of an Internet Routing Protocol*. Addison-Wesley, 1998.

[61] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, "IDMaps : A global Internet host distance estimation service," *IEEE/ACM Transactions on Networking*, vol. 9, no. 5, pp. 525–540, Oct. 2001.

[62] T. S. E. Ng and H. Zhang, "Predicting Internet network distance with coordinates-based approaches," in *Proc. IEEE INFOCOM*, New York, NY, USA, June 2002.

[63] K. P. Gummadi, S. Saroiu, and S. D. Gribble, "King : Estimating latency between arbitrary Internet end hosts," in *ACM Internet Measurement Workshop*, Marseille, France, Nov. 2002.

[64] M. Pias, J. Crowcroft, S. Wilbur, T. Harris, and S. Bhatti, "Lighthouses for scalable distributed location," in *Proc. the Second International Workshop on Peer-to-Peer Systems - IPTPS*, Berkeley, CA, USA, Feb. 2003.

[65] Y. Shavitt and T. Tankel, "Big-bang simulation for embedding network distances in Euclidean space," in *Proc. IEEE INFOCOM*, San Francisco, CA, USA, Mar. 2003.

[66] L. Tang and M. Crovella, "Virtual landmarks for the Internet," in *ACM Internet Measurement Conference*, Miami, FL, USA, Oct. 2003.

[67] H. Lim, J. C. Hou, and C. H. Choi, "Constructing Internet coordinate system based on delay measurement," in *ACM Internet Measurement Conference – IMC*, Miami, FL, USA, Oct. 2003.

[68] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, "Vivaldi : A decentralized network coordinate system," in *Proc. ACM SIGCOMM*, Portland, OR, USA, Aug. 2004.

[69] G. Ballintijn, M. van Steen, and A. S. Tanenbaum, "Characterizing Internet performance to support wide-area application development," *Operating Systems Review*, vol. 34, no. 4, pp. 41–47, Oct. 2000.

[70] A. Ziviani, "Qualité de service et localisation d'hôtes (quality of service and location-awareness)," Ph.D. dissertation, Université Pierre et Marie Curie (Paris 6), Paris, France, Dec. 2003.

[71] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Improving the accuracy of measurement-based geographic location of Internet hosts," *Computer Networks, Elsevier Science*, vol. 47, no. 4, pp. 503–523, Mar. 2005.

[72] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. the International World Wide Web Conference - WWW10*, Hong Kong, May 2001.

[73] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Demographic placement for Internet host location," in *Proc. IEEE GLOBECOM*, San Francisco, CA, USA, Dec. 2003.

[74] ——, "Placement issues in measurement-based host location," in *Proc. the 9th European Summer School and IFIP Workshop on Next Generation Networks - EUNICE*, Budapest - Balatonfüred, Hungary, Sept. 2003, pp. 173–180.

[75] T. Brinkhoff, *City Population*, http://www.citypopulation.de.

[76] *The World Factbook 2002*, Central Intelligence Agency (CIA), Jan. 2002, http://www.cia.gov/cia/publications/factbook.

[77] C. Toregas, R. Swain, C. Revelle, and L. Bergman, "The location of emergency service facilities," *Operations Research*, vol. 19, no. 6, pp. 1363–1373, 1971.

[78] M. S. Daskin, *Network and Discrete Location*. New York, NY : John Wiley & Sons, 1995.

[79] J. Current, M. Daskin, and D. Schilling, *Discrete Network Location Models*. Springer-Verlag, 2002, in Facility Location, pp. 81–118.

[80] B. Gueye, A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Two-tier geographic location of internet hosts," in *Proc. IEEE International Conference on High Speed Networks and Multimedia Communications - HSNMC*, ser. Lecture Notes in Computer Science (LNCS), Toulouse, France, June 2004.

[81] B. Wong, I. Stoyanov, and E. Sirer, "Geolocalization on the internet through constraint satisfaction," in *Proc. Workshop on Real, Large Distributed Systems – WORLDS*, Seattle, Washington, USA, Nov. 2006.

[82] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP topologies with rocketfuel," in *Proc. ACM SIGCOMM*, Pittsburgh, PA, Aug. 2002.

[83] R. Percacci and A. Vespignani, "Scale-free behavior of the Internet global performance," *The European Physical Journal B - Condensed Matter*, vol. 32, no. 4, pp. 411–414, Apr. 2003.

[84] C. J. Bovy, H. T. Mertodimedjo, G. Hooghiemstra, H. Uijterwaal, and P. van Mieghem, "Analysis of end-to-end delay measurements in Internet," in *Proc. PAM workshop*, Fort Collins, CO, USA, Mar. 2002.

[85] S. van Langen, X. Zhou, and P. van Mieghem, "On the estimation of Internet distances using landmarks," in *Proc. the International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking – NEW2AN*, St. Petersburg, Russia, Feb. 2004.

[86] S. Moon, P. Skelly, and D. Towsley, "Estimation and removal of clock skew from network delay measurements," in *Proc. IEEE INFOCOM*, Piscataway, NJ, USA, Mar. 1999, pp. 227–234.

[87] V. Chvatal, *Linear Programming.* New York, NY : W. H. Freeman, 1983.

[88] J. Venn, "On the diagrammatic and mechanical representation of propositions and reasonings," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 9, 1880.

[89] *NLANR Active Measurement Project*, 1998, http://watt.nlanr.net/.

[90] *RIPE Test Traffic Measurements*, 2000, http://www.ripe.net/ttm/.

[91] *GeoLIM Project*, 2005, http://planetlab-01.ipv6.lip6.fr:10000/cbg.php/.

[92] *PlanetLab : An open platform for developing, deploying, and accessing planetary-scale services*, 2002, http://www.planet-lab.org.

[93] L. Paterson and T. Roscoe, "The Design Principles of PlanetLab," *Operating Systems Review*, vol. 40, no. 1, pp. 11–16, January 2006.

[94] *R Core Team*, 1997, http://www.r-project.org/.

[95] T. Ylonen, T. Kivinen, and M. Saarinen, "Ssh protocol archtecture," *Internet Draft IETF*, Nov. 1997.

[96] D. Krioukov, K. Fall, and X. Yang, "Compact routing on Internet-like graphs," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004.

[97] N. Spring, R. Mahajan, and T. Anderson, "Quantifying the causes of path inflation," in *Proc. ACM SIGCOMM*, Karlsruhe, Germany, Aug. 2003.

[98] L. Gao and F. Wang, "The extent of as path inflation by routing policies," in *IEEE Global Internet Symposium*, 2002.

[99] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The end-to-end effects of Internet path selection," in *Proc. ACM SIGCOMM*, Cambridge, MA, USA, Sept. 1999.

[100] H. Tangmunarunkit, R. Govindan, and S. Shenker, "Internet path inflation due to policy routing," in *SPIE ITCom*, 2001.

[101] H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin, "The impact of routing policy on internet paths," in *Proc. IEEE INFOCOM*, Anchorage, AK, USA, Apr. 2001.

[102] H. Zheng, E. K. Lua, M. Pias, and T. Griffin, "Internet Routing Policies and Round-Trip-Times," in *Proc. Passive and Active Measurement Workshop – PAM*, Boston, MA, USA, Apr. 2005.

[103] S. Keshav, "A control-theoretic approach to flow control," in *Proc. ACM SIGCOMM*, Zurich, Switzerland, Sept. 1991, pp. 3–15.

[104] J. C. R. Bennett and H. Zhang, "WF2q : Worst-case fair weighted fair queueing," in *Proc. IEEE INFOCOM*, 1996, pp. 120–128.

[105] L. Georgiadis, R. Guérin, V. Peris, and K. N. Sivarajan, "Efficient network QoS provisioning based on per node traffic shaping," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 482–501, 1996.

[106] J. Postel, *Internet Control Message Protocol*, Sept. 1981, rFC 792 IETF.

[107] N. Spring, D. Wetherall, and T. Anderson, "Scriptroute : A facility for distributed internet measurement flash," in *Proc. of USENIX USITS'03*, Mar. 2003.

[108] *MaxMind*, MaxMind LLC, http://www.maxmind.com.

[109] *Ed Williams*, http://http://williams.best.vwh.net/avform.html.

[110] *Fédération Aéronotique Internationale*, http://http://www.fai.org.

[111] "Geographic location/privacy (geopriv) IETF working group," 2003, http://www.ietf.org/html.charters/geopriv-charter.html.

Glossaire

AfricNIC African Network Information Centre, 12
AMP Active Measurement Project, 43
AP Adresse Préfixe, 15
APNIC Asia Pacific Network Information Centre, 12
ARIN American Registry for Internet Numbers, 12

BBS Bing-Bang Simulation, 16
BGP Border Gateway Protocol, 15

DNS Domaine Name Server, 11
DSR Dynamic Source Routing, 9

eBGP Exterior Border Gateway Protocol, 16

FAI Fournisseur d'Accès Internet, 4
FIFO First In First Out, 60

GeoBuD **Geo**location using **Bu**ffering **D**elay estimation, 61
GeoLIM Geographic Location of Internet hosts with Multilateration, 41
geopriv Geographic Location/Privacy, 82
GGAR GeoCast - Geographic Addressing and Routing, 9
GLONASS GLObal'naya NAvigatsionaya Sputnikova Sputnikovaya Sistema, 7
GNP Global Network Positioning, 16
GNSS Global Navigation Satellite System, 31
GPS Global Positioning System, 4
GPSR Greedy Perimeter Stateless Routing for Wireless Networks, 11

GRP Geographical Routing, 11
GSM Global System for Mobile Communication, 8

HLR Home Location Register, 8

ICMP Internet Control Message Protocol, 62
ICS Internet Coordinate System, 16
IETF Internet Engineering Task Force, 82
IMSI International Mobile Subscriber Identity, 8
IP Internet Protocol, 2

LACNIC Latin American and Caribbean Internet Addresses Registry, 12
LAR Location Aided Routing, 9
LIP6 Laboratoire d'Informatique de Paris6, 13

MANET Mobile Ad hoc NETwork, 9
MCLM Maximum Covering Location Model, 24

NLANR National Laboratory for Applied Network Research, 41

ping Packet Internet Groper, 19

RADAR RAdio Detection And Ranging, 8
RFC Request For Comments, 11
RIPE NCC RIPE Network Coordination Centre, 12
RIR Regional Internet Registries, 11
RTT Round Trip Time, 20

SimPA Simple Positioning Algorithm, 4

TCP Transmission Control Protocol, 12
TTL Time To Live, 63
TTM Test Traffic Measurements, 43

UCL Université Catholique de Louvain, 13
UDP User Datagram Protocol, 62

VLR Visitor Location Register, 8

WFQ Weighted Fair Queuing, 60

Table des figures

Liste des tableaux