
Vers la Localisation Géographique d'Hôtes dans l'Internet basée sur la Multilatération

Bamba Gueye* — **Artur Ziviani**** — **Mark Crovella***** — **Serge Fdida***

* *Université Pierre et Marie Curie
Laboratoire d'Informatique de Paris 6
8, rue du Capitaine Scott
75015 Paris France
{bamba.gueye, serge.fdida}@lip6.fr*

** *Laboratório Nacional de Computação Científica (LNCC)
Av. Getúlio Vargas, 333
25651-075 - Petrópolis - RJ - Brasil
ziviani@lncc.br*

*** *Department of Computer Science
Boston University
111 Cummington St - 02215
Boston, MA - USA
crovella@cs.bu.edu*

RÉSUMÉ. L'inférence de la localisation géographique d'hôtes dans l'Internet permet l'émergence de nouvelles applications très variées. Jusqu'à présent, la localisation d'un hôte cible est fournie par la position des hôtes références, hôtes dont on connaît les positions géographiques. Ainsi le nombre d'endroits possibles où on peut localiser un hôte cible est égal au nombre d'hôtes références, conduisant ainsi à un espace discret de réponses. Nous proposons une technique de Localisation Géographique basée sur la Multilatération (LGM) 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 contrairement aux approches précédentes. LGM transforme les mesures de délai en distance géographique surestimée, malgré les délais supplémentaires dus aux congestions, et la non linéarité des chemins entre les hôtes. LGM utilise la multilatération avec ces distances géographiques surestimées pour inférer la position de l'hôte. Les résultats obtenus montrent que LGM est plus performante que les précédentes techniques de localisation et, de surcroît, est capable d'attribuer une zone de confiance à chaque hôte localisé.

ABSTRACT. Geolocation of Internet hosts enables a diverse and interesting new class of location-aware applications. Previous measurement-based approaches use reference hosts, called land-

marks, 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 a geographic location based on multilateration, which infers the geographic location of Internet hosts. Multilateration allows to establish a continuous space of answers instead of a discrete one. CBG accurately transforms delay measurements to geographic distance constraints in spite of queuing delays and the absence of great-circle paths between hosts, and then uses multilateration to infer the geolocation of the target host. Our experimental results show that CBG outperforms the previous measurement-based geolocation techniques. Moreover, in contrast to previous approaches, our method is able to assign a confidence region to each given location estimate.

MOTS-CLÉS : Localisation, multilatération, mesures de délai

KEYWORDS: Geolocation, multilateration, delay measurements

1. Introduction

Avec le développement des nouvelles technologies de l'information, des nouveaux services ont fait leur apparition, notamment des services dits de "proximité" basés sur la localisation des clients. Nous pouvons citer comme exemple la publicité ciblée, la sélection automatique de la langue à la connexion, la diffusion de contenu suivant une politique géographique, et l'acceptation d'une transaction bancaire seulement à partir d'un endroit pré-établi. Les techniques de localisation basées sur des mesures tentent de déterminer la position géographique d'un hôte en se basant sur la connaissance de son adresse IP. C'est ainsi que [ZIV 04b, PAD 01] proposent d'utiliser la position de l'hôte référence, hôte dont on connaît la position géographique, le plus proche en terme de délai comme possible localisation de l'hôte cible. 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. Cette approche manque de précision car l'hôte référence le plus proche, qui fournit sa position comme estimation de localisation de la cible, n'est pas forcément proche d'elle.

Nous proposons une nouvelle technique de Localisation Géographique basée sur la Multilatération (LGM) pour résoudre ce problème. 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. Connaissant la distance géographique *surestimée* entre la cible et chaque hôte référence, LGM fournit à l'instar du système de positionnement par satellites (GPS) [ENG 99] une estimation de localisation. 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 nécessairement corrélée au délai [BAL 00]. Cela est dû aux congestions qui existent dans les réseaux, ajoutant un délai supplémentaire dans les mesures, et à la non linéarité des chemins entre les hôtes [SUB 02].

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 [PER 03], à partir de toute mesure de délai faite entre deux points, nous pouvons calculer leur distance *maximale*. Cette distance maximale 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.

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. LGM vise à le faire grâce à un auto-calibrage des mesures de délai obtenues entre les hôtes références (voir section 3). Après l'auto-calibrage, les mesures de délai entre les hôtes références et l'hôte cible, sont transformées en des distances appelées distances géographiques *surestimées*, car formées par les distances géographiques

réelles et les distances induites par les délais supplémentaires qui s'ajoutent aux mesures. LGM utilise ces distances géographiques surestimées pour inférer la zone géographique dans laquelle se trouve l'hôte cible. LGM considère le centre de cette zone comme la localisation de l'hôte cible. Contrairement aux précédentes techniques de localisation géographique, LGM est capable de fournir une zone de confiance pour chaque hôte localisé. Cela permet aux applications qui l'utilisent d'évaluer la fiabilité de l'estimation par rapport à leurs exigences.

Pour l'évaluation de LGM nous avons utilisé les mesures de délai d'hôtes localisés à travers les États Unis et l'Europe de l'Ouest. Les résultats montrent que LGM est plus performante en précision que les précédentes techniques de localisation géographique, basées sur des mesures. Ainsi l'erreur médiane de distance obtenue est inférieure à 25 km pour l'ensemble des hôtes localisés en Europe de l'Ouest, et 100 km pour ceux localisés aux États Unis. Nous avons remarqué que dans la plupart des cas, la zone de confiance que LGM fournit est raisonnable, car assimilable à la superficie d'un petit pays comme la Belgique en Europe ou un petit état comme le Maryland aux États Unis.

Ce papier est organisé comme suit. La section 2 décrit les principales motivations qui ont attiré à la localisation d'un hôte, dresse l'état de l'art du domaine et montre les contributions que LGM a apportées. Dans la section 3 nous présentons la technique LGM et ses différentes caractéristiques. La section 4 illustre les différents résultats obtenus en appliquant LGM. Enfin la section 5 conclut notre travail et présente quelques perspectives pour le long terme.

2. Techniques de localisation géographique

2.1. Motivations

La localisation des hôtes permet d'avoir une dimension géographique de l'Internet mais elle est beaucoup plus motivée par des applications commerciales. Ces applications peuvent exiger différentes échelles de localisation, à savoir l'échelle d'un pays, d'une région ou d'une ville. Les techniques de localisation permettent ainsi de proposer des services personnalisés tels que :

– *La publicité ciblée* : les internautes peuvent avoir différentes préférences suivant l'endroit où ils résident. Ainsi les annonceurs pourront définir plusieurs stratégies de marketing.

– *La diffusion restreinte de contenu* : si nous avons une politique régionale de diffusion, une application utilisant la localisation géographique peut déterminer quels sont les clients qui ont le droit d'accéder au contenu diffusé. Toutefois un organisme de régulation est obligatoire.

– *Identification basée sur la localisation du client* : avec la croissance du commerce en ligne, des endroits pré-établis, à partir desquels on peut accepter des transactions, peuvent être établis pour un client donné. Ainsi une transaction établie à partir

d'un endroit quelconque peut entraîner un refus pour non respect des clauses.

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

2.2. *État de l'art*

La RFC 1876 [DAV 96] propose d'ajouter des informations de localisation dans les noms DNS (Domaine Name Server). Cependant cette proposition ne fut pas largement adoptée, car les administrateurs n'étaient pas trop motivés pour ajouter des enregistrements de localisation dans les bases de données DNS. C'est à travers des bases de données Whois que des outils comme IP2LL [Uni] et NetGeo [MOO 00] tentent d'inférer la localisation d'un hôte à partir de son adresse IP. Ces informations peuvent être désuètes et inexactes.

Padmanabhan et Subramanian [PAD 01] quant à eux ont développé trois techniques pour inférer la localisation géographique d'un hôte. La première technique GeoTrack est basée sur l'utilisation de la commande Traceroute [V.] qui permet de déterminer un chemin depuis un serveur de mesure vers l'hôte qu'on veut localiser. GeoTrack déduit la localisation d'un hôte à partir des noms fournis par le DNS de l'hôte cible ou des routeurs qui lui sont proches. Elle est similaire à d'autres techniques comme VisualRoute [Vis], GTrace [CAI], et à la méthode utilisée par le projet Sarang-World Traceroute [Sar]. Les noms DNS dans l'Internet contiennent parfois certaines indications sur la localisation. Par exemple le nom `bcr1-so-2-0-0.Paris.cw.net` indique un routeur localisé à Paris (France). Toutefois 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.

La deuxième technique, GeoCluster, se base sur la notion de cluster [KRI 00] 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. Pour inférer la localisation d'un hôte, GeoCluster détermine d'abord son cluster géographique en utilisant une base de données contenant une association d'adresses IP et leurs localisations. Connaissant la localisation de quelques hôtes qui s'y trouvent, elle déduit la localisation du cluster en entier. 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 sont fournies par les utilisateurs qui sont néanmoins peu fiables.

La troisième technique, GeoPing, est la plus proche de LGM, et exploite une possible corrélation entre délai et distance géographique. L'hypothèse de base de GeoPing est que des hôtes ayant un délai similaire par rapport à d'autres hôtes fixes (des serveurs sondes par exemple) tendent à être situés dans une même zone géographique. Cette technique s'inspire des techniques utilisées dans RADAR [BAH 00] 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 à l'hôte dont on veut déterminer sa localisation. Le nombre d'endroits possibles, où on peut localiser l'hôte cible, est alors limité au nombre d'hôtes références, d'où un espace discret de réponses. Par conséquent, le nombre et le placement des hôtes références jouent un rôle important dans la précision de l'estimation de localisation [ZIV 04a]. L'amélioration de la technique GeoPing passe par une augmentation du nombre d'hôtes références. Dans la section 4 nous comparons LGM à l'approche DNS et à la technique GeoPing.

2.3. Contributions

LGM est la première technique dans le domaine de la localisation à utiliser la multilatération pour inférer la position d'un hôte. Ses principales contributions sont :

- LGM établit une relation dynamique entre les adresses IP et leur localisation géographique grâce à des mesures de délai faites périodiquement entre les hôtes références. Nous obtenons un système distribué où les hôtes références font un auto-calibrage permettant à la technique LGM de s'adapter aux différents états du réseau contrairement aux techniques précédentes qui se basent sur des données statiques.

- L'apport principal de la technique LGM est sa capacité à transformer les mesures de délai en distances géographiques surestimées, utilisées par la multilatération. Ainsi, en utilisant la multilatération, nous obtenons un espace continu d'endroits où on peut localiser un hôte contrairement aux autres techniques de localisation basées sur des mesures de délai.

- LGM fournit également une zone de confiance pour chaque hôte localisé offrant aux applications qui l'utilisent la possibilité d'évaluer la précision de l'estimation par rapport à leurs exigences.

3. Localisation Géographique basée sur la Multilatération

3.1. 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 leur localisation. Lorsqu'on considère des distances, ce procédé est appelé multilatération. Par contre, si c'est des angles, il est appelé 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, en étant celui le plus approprié, est utilisé dans le reste de cet article au détriment du terme triangulation, malgré sa popularité.

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) [ENG 99] pour localiser un récepteur GPS utilise la multilatération. On mesure la distance entre le récepteur et trois satellites dont leurs 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 exact. Par contre, la transformation des mesures de délai en distance géographique est un véritable challenge dans l'Internet. C'est la raison pour laquelle la multilatération est restée inutilisée. Par la suite nous expliquons les principaux aspects de LGM et comment se fait la transformation des mesures de délai en distances géographiques surestimées.

Soit un ensemble $\mathcal{H} = \{H_1, H_2, \dots, H_K\}$ de K hôtes références. Connaissant les délais entre un hôte cible et ces hôtes références, notre principal objectif est d'estimer les distances géographiques correspondantes. En effet, le délai de bout en bout est la somme des délais de propagation et de transmission, et des délais d'attente dans les files des routeurs [BOV 02]. Au délai effectif mesuré s'ajoute donc un délai supplémentaire induit par ces distorsions. Par conséquent, l'estimation de distance fournie par LGM est composée de la distance géographique réelle à laquelle s'ajoute une distance induite par ces distorsions. Ainsi, l'estimation de distance fournie par LGM est appelée par définition distance géographique surestimée, car étant la somme de la distance géographique réelle et de la distance induite par les distorsions.

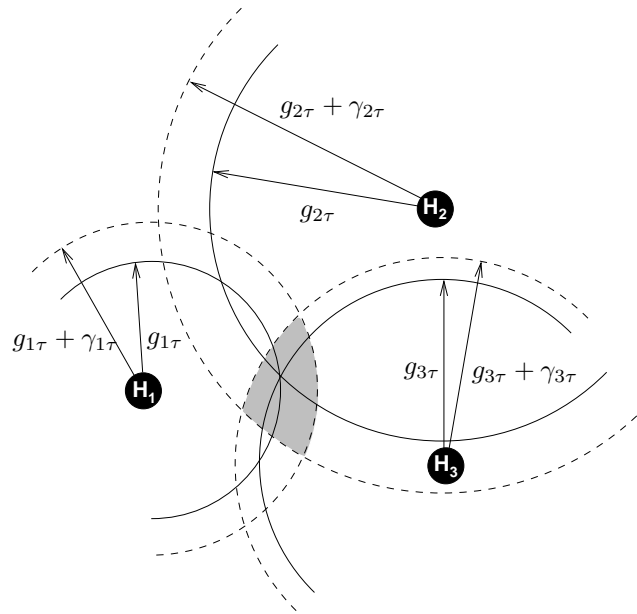


Figure 1. Multilatération utilisant des distances géographiques surestimées.

La Figure 1 montre le principe de la multilatération où on utilise un ensemble $\mathcal{H} = \{H_1, H_2, H_3\}$ d'hôtes références. Chaque hôte référence tente d'inférer sa distance géographique surestimée par rapport à l'hôte cible τ . Cette distance est $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$, où $g_{i\tau}$ représente la distance géographique réelle et $\gamma_{i\tau}$ une distance additive. Cette distance additive $\gamma_{i\tau}$ provient du délai supplémentaire, dû aux distorsions et, imbriqué dans le délai de bout en bout. La localisation de l'hôte cible se trouve quelque part à l'intérieur de la zone grise (voir Figure 1). Cette zone 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.

3.2. Transformation des mesures de délai en distances géographiques surestimées

Avant d'introduire comment LGM convertit les mesures de délai en distances géographiques surestimées, regardons d'abord la relation pouvant exister entre distance géographique et délai. La Figure 2 illustre un exemple choisi parmi les résultats décrits dans la section 4. L'axe des abscisses représente la distance géographique réelle et l'axe des ordonnées le délai mesuré entre un hôte référence H_i et les autres hôtes références restants. Les définitions de "droite optimale" et "droite théorique" dans la Figure 2 seront expliquées par la suite.

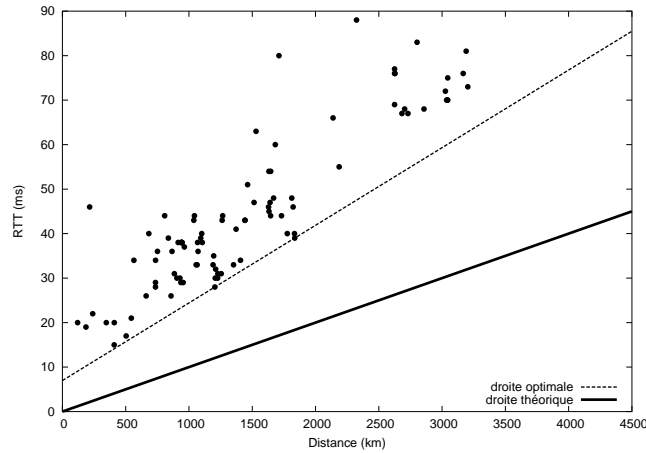


Figure 2. Exemple montrant la relation entre distance géographique et délai.

[ZIV 04b, PAD 01, van 04] 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 2, nous pensons que cette droite ne traduit pas au mieux la relation entre dis-

tance 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 [GUE 04]. 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 H_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 $y = mx + b$, 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, alors 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 "droite théorique" montrée dans la Figure 2. Si nous avons 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.

Ainsi pour modéliser la relation entre délai et distance géographique, nous définissons une "droite optimale" pour chaque hôte référence H_i comme la droite $y = m_i x + b_i$ qui est la plus proche, mais en dessous de tous les points (x, y) et dont l'ordonnée à l'origine i.e. b_i n'est pas négative. Car considérer un délai négatif serait un non sens. La droite optimale est considérée comme la droite qui prend compte de la distorsion la plus petite soit elle entre le délai et la distance géographique. L'ordonnée à l'origine de cette droite traduit la présence d'une source de distorsion. Chaque hôte référence calcule sa propre droite optimale par rapport à tous les autres hôtes références. Soit un hôte référence H_i , nous calculons le délai d_{ij} et la distance géographique g_{ij} vers chaque hôte référence H_j , où $i \neq j$. Nous cherchons alors pour chaque hôte référence H_i la pente m_i et l'ordonnée à l'origine b_i qui déterminent la droite optimale $y = m_i x + b_i$. La droite optimale de chaque hôte référence H_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

$$y - m_i x - b_i \geq 0, \quad \forall i \neq j, \quad (1)$$

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

$$\min_{\substack{b_i \geq 0 \\ m_i \geq m}} \left(\sum_{i \neq j} y - m_i x - b_i \right), \quad (2)$$

où m représente la pente de la droite théorique. L'équation 1 détermine la droite optimale de chaque hôte référence H_i . Quant à l'équation 2, elle est utilisée pour trouver m_i et b_i .

Chaque hôte référence utilise sa propre droite optimale 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 H_i et l'hôte cible τ , est obtenue à partir de l'équation ci-dessous

$$\hat{g}_{i\tau} = \frac{d_{i\tau} - b_i}{m_i}. \quad (3)$$

où $d_{i\tau}$, m_i , et b_i représentent respectivement le délai de H_i vers τ , la pente et l'ordonnée à l'origine de la droite optimale de l'hôte référence H_i .

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

3.3. Applications de la multilatération

La technique LGM utilise une approche géométrique pour estimer la localisation d'un hôte cible τ . Chaque hôte référence H_i infère la distance géographique surestimée $\hat{g}_{i\tau} = g_{i\tau} + \gamma_{i\tau}$ qui la sépare à l'hôte cible τ en utilisant l'équation 3. Ainsi chaque hôte référence H_i estime que l'hôte cible τ se trouve quelque part à l'intérieur du cercle $\mathcal{C}_{i\tau}$ centré en H_i et de rayon $\hat{g}_{i\tau}$ (analogue à l'exemple de la Figure 1). Par exemple si nous avons K hôtes références, nous avons $\mathbf{C}_\tau = \{\mathcal{C}_{1\tau}, \mathcal{C}_{2\tau}, \dots, \mathcal{C}_{K\tau}\}$ cercles qui constituent un diagramme de Venn 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} donnée par

$$\mathcal{R} = \bigcap_i^K \mathcal{C}_{i\tau}. \quad (4)$$

Cette région \mathcal{R} correspond à la zone grise de la Figure 1 et contient la position réelle de l'hôte cible. Il faut noter que la région \mathcal{R} est convexe car les cercles $\mathcal{C}_{i\tau}$ sont convexes chacun, et par définition l'intersection d'un ensemble de convexes est convexe.

4. Résultats

4.1. Paramètres expérimentaux

Pour pouvoir évaluer la précision de LGM, nous n'avons besoin que des données fournies par des hôtes dont on connaît leur localisation géographique. Cette exigence limite le nombre d'hôtes qu'on peut utiliser. Ainsi, pour nos expériences nous avons utilisé des hôtes de deux ensembles :

- RIPE : les données ont été collectées à partir du réseau européen RIPE [RIP] grâce 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).

- NLANR AMP : les données ont été collectées grâce à l'organisation NLANR [NLA] 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 (E.U) 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.

Nous construisons deux matrices de délai \mathbf{D}_{ripe} de dimension (42×42) et \mathbf{D}_{amp} de dimension (95×95) à partir des mesures obtenues dans chaque ensemble. Chaque hôte est considéré comme hôte référence, nous obtenons ainsi deux ensembles d'hôtes références : $\mathcal{H}_{\text{ripe}} = \{H_1, H_2, \dots, H_{42}\}$ et $\mathcal{H}_{\text{amp}} = \{H_1, H_2, \dots, H_{95}\}$. 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 droite optimale de l'hôte référence choisi comme hôte cible n'est pas utilisée lors de sa localisation c'est seulement les droites optimales des autres hôtes références restants qui sont utilisées. Ce processus est répété pour la localisation de chaque hôte référence pris comme cible dans chaque ensemble.

4.2. Obtention de l'estimation de localisation de l'hôte cible

A partir des distances géographiques surestimées, LGM détermine pour chaque hôte cible τ l'ensemble des cercles $\mathbf{C}_\tau = \{C_{1\tau}, C_{2\tau}, \dots, C_{K\tau}\}$ (voir section 3.3) où $K=95$ pour les E.U et $K=42$ pour l'E.O . Chaque cercle de \mathbf{C}_τ a pour centre son hôte référence correspondant et pour rayon la distance géographique surestimée $\hat{g}_{i\tau}$. La Figure 3 montre un exemple extrait à partir de nos résultats obtenus et illustre la

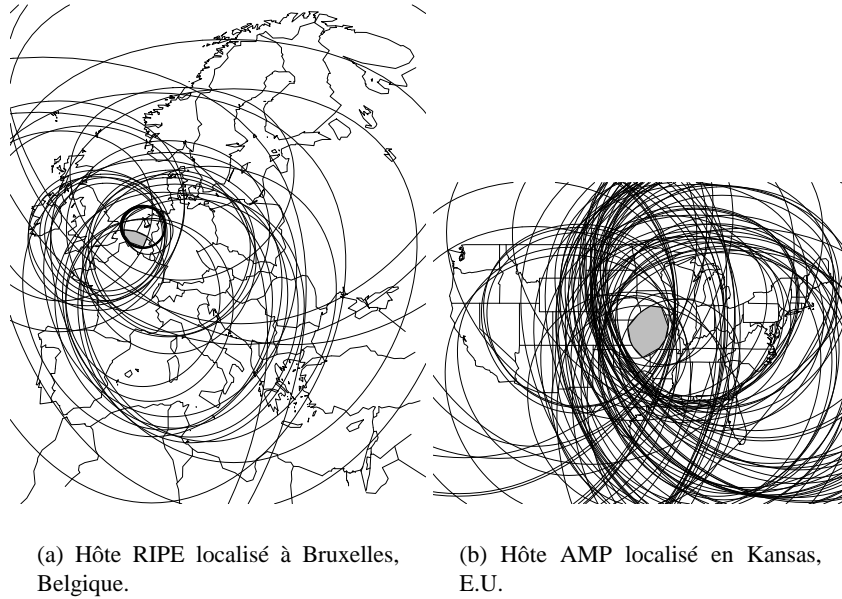


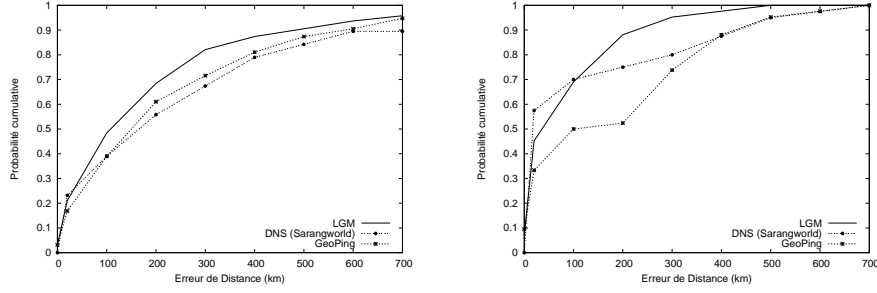
Figure 3. Exemple de zone d'estimation de localisation de deux hôtes cibles.

méthodologie de LGM. La Figure 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'E.O. De la même manière la Figure 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 dans les Figures 3(a) et 3(b) représente la zone d'intersection \mathcal{R} de ces cercles. Cette région \mathcal{R} est la zone de confiance que LGM 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 méthode LGM que l'exemple de la Figure 3 a été choisie (la section 4.4 répertorie la taille des différentes zones de confiance obtenues).

4.3. Processus de localisation d'un hôte cible

La région \mathcal{R} représente l'estimation de localisation fournie par LGM 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 pas déterminer la forme géométrique correspondante à cette région, d'où trouver son centre nous est impossible. Ainsi nous avons cherché à approximer cette région \mathcal{R} par un polygone.



(a) Hôtes situés aux États Unis.

(b) Hôte situés en Europe de l'Ouest.

Figure 4. Erreur de distance de LGM, de la méthode DNS et de GeoPing.

Le polygone obtenu fournit l'estimation de localisation de l'hôte cible, et sa surface la zone de confiance associée à l'estimation. 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} , car elle est convexe. Par exemple, dans la Figure 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 $v_0 = (x_0, y_0), \dots, v_{N-1} = (x_{N-1}, y_{N-1})$ est donnée par

$$A = \frac{1}{2} \sum_{n=0}^{N-1} \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix} \quad (5)$$

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

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

et

$$c_y = \frac{1}{6A} \sum_{n=0}^{N-1} (y_n + y_{n+1}) \begin{vmatrix} x_n & x_{n+1} \\ y_n & y_{n+1} \end{vmatrix}. \quad (7)$$

Le point de coordonnées (c_x, c_y) et la surface \mathcal{A} représentent respectivement l'estimation de localisation et la zone de confiance de l'hôte cible. 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 une méthode basée sur les noms DNS (*voir* le projet SarangWorld Traceroute [Sar]) et par GeoPing qui utilise un espace discret de réponses [ZIV 04b, PAD 01]. La Figure 4 montre la fonction de probabilité cumulative de l'erreur de distance obtenue en utilisant LGM, la méthode basée sur le DNS et GeoPing. LGM 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 démontré que, si nous avons un espace discret d'endroits où on peut localiser un hôte, le nombre d'hôtes références et leurs placements jouent un rôle considérable dans la précision de l'estimation [ZIV 04a].

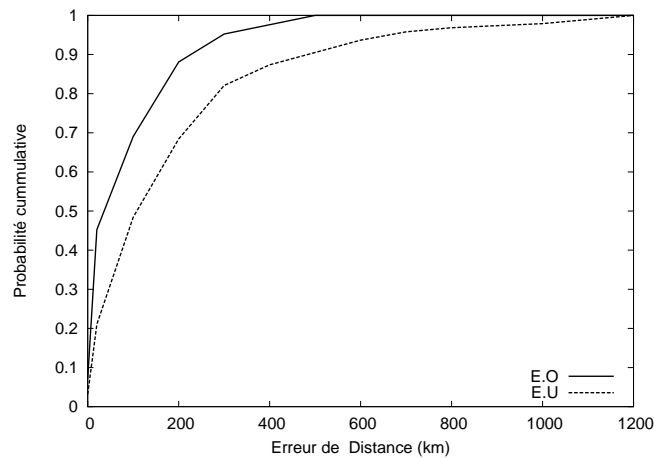


Figure 5. Erreur de distance de LGM aux E.U et en E.O.

Dans la Figure 5, nous comparons les erreurs de distance obtenues pour localiser des hôtes situés aux E.U et en E.O en utilisant LGM. L'erreur moyenne de distance est évaluée à 182 km au niveau des E.U alors qu'elle est de 78 km pour l'E.O. La plupart des hôtes ont une assez bonne estimation de localisation au niveau de nos deux ensembles. L'erreur médiane de distance est de 95 km et 80 % des hôtes sont localisés avec une erreur inférieure à 277 km. Pour l'ensemble des hôtes situés en E.O, l'erreur médiane est de 22 km et 80 % sont localisés avec une erreur inférieure à 134 km.

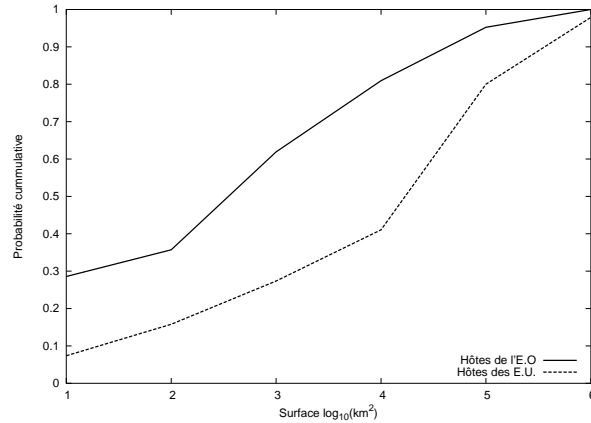


Figure 6. Zone de confiance fournie par LGM en km^2 .

4.4. Zone de confiance de l'estimation de localisation

La région \mathcal{R} représente la zone de confiance, évaluée en km^2 , que LGM associe à l'estimation de localisation. Plus la surface \mathcal{R} est petite plus l'estimation de localisation est précise. Ainsi la force de LGM est sa capacité à fournir une zone de confiance, contrairement aux précédentes techniques 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 6 illustre la fonction de probabilité cumulative des zones de confiance associées à l'estimation de localisation des hôtes références situés aux E.U et en E.O. Les résultats montrent que pour les E.U, 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 surface d'un pays comme le Portugal ou un état des États Unis comme Indiana. Pour les hôtes situés en E.O, 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 E.U sont localisés avec une zone de confiance inférieure à 10^3 km^2 alors qu'en Europe de l'Ouest 65 % des hôtes le sont. Cette surface est équivalente à la taille d'une grande région métropolitaine.

5. Conclusion

Cet article a proposé une nouvelle technique de localisation géographique basée sur la multilatération (LGM) pour inférer la position d'un hôte cible à partir de mesures de délai. Transformer les mesures de délai en distance géographique surestimée avec précision est un challenge en raison de beaucoup de particularités inhérentes à l'utilisation d'Internet. La non linéarité du chemin entre deux hôtes et la congestion dans les réseaux pouvant occasionner des délais supplémentaires dans les mesures en

sont des exemples. Cependant LGM montre que la transformation des mesures de délai en distance géographique surestimée est possible. De plus, LGM montre que la surestimation de cette distance géographique est souvent assez petite permettant ainsi de localiser avec précision l'hôte cible en utilisant la multilatération. LGM établit une relation dynamique entre délai et distance géographique grâce à un auto-calibrage des hôtes références en utilisant leur droite optimale.

Les résultats obtenus montrent que LGM est plus performante que les précédentes techniques de localisation. Ainsi, l'erreur médiane de distance obtenue pour les hôtes références localisés aux E.U 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 E.U et 100 km pour l'Europe Occidentale. En outre, contrairement aux méthodes précédentes, LGM fournit une zone de confiance pour chaque estimation de localisation. Cela est important pour permettre aux applications d'évaluer si la précision fournie est suffisante. La plupart des hôtes ont été localisés avec une zone de confiance assez petite à l'ordre d'une région métropolitaine, d'où une précision de notre estimation de localisation. Pour accroître la précision de l'estimation de l'hôte cible, nous envisageons dans nos travaux futurs d'utiliser des hôtes références régionaux après avoir déterminé la zone de confiance de l'hôte cible.

Les mesures de délai exploitées ont été obtenues à partir de réseaux fortement connectés et géographiquement adjacents. Cependant notre travail présente un intérêt du fait que la connectivité dans les réseaux s'est fortement améliorée durant cette dernière décennie. Il faut noter que la relation entre délai et distance géographique est très forte dans ces régions [ZIV 04b, CLA 02, YOO 02]. La localisation à partir ou vers des hôtes situés un peu partout dans l'Internet, par exemple PlanetLab [pla], est envisagé pour nos travaux à long terme. De même que l'utilisation d'une base de données, où on enregistre les couples IP-Localisation des hôtes déjà localisés, afin d'éviter des mesures répétitives.

6. Bibliographie

- [BAH 00] BAHL P., PADMANABHAN V. N., « RADAR : An In-Building RF-Based User Location and Tracking System », *Proc. of the IEEE INFOCOM'2000*, Tel-Aviv, Israel, mars 2000.
- [BAL 00] BALLINTIJN G., VAN STEEN M., TANENBAUM A. S., « Characterizing Internet Performance to Support Wide-Area Application Development », *Operating Systems Review*, vol. 34, n° 4, 2000, p. 41–47.
- [BOV 02] BOVY C. J., MERTODIMEDJO H. T., HOOGHIEMSTRA G., UIJTERWAAL H., VAN MIEGHEM P., « Analysis of End-to-end Delay Measurements in Internet », *Proc. of the Passive and Active Measurement Workshop - PAM'2002*, Fort Collins, CO, USA, mars 2002.
- [CAI] CAIDA, « GTrace », <http://www.caida.org/tools/visualization/gtrace/>.
- [CLA 02] CLAFFY K., « Internet measurement : myths about Internet data », Talk at NANOG24 Meeting, février 2002, <http://www.caida.org/outreach/presentations/>

Myths2002/.

- [DAV 96] DAVIS C., VIXIE P., GOOWIN T., DICKINSON I., « A Means for Expressing Location Information in the Domain Name System », *Internet RFC 1876*, , 1996.
- [ENG 99] ENGE P., MISRA P., « Special Issue on Global Positioning System », *Proceedings of the IEEE*, vol. 87, n° 1, 1999, p. 3–15.
- [GUE 04] GUEYE B., ZIVIANI A., CROVELLA M., FDIDA S., « Constraint-Based Geolocation of Internet Hosts », *Proc. of the ACM Sigcomm Internet Measurement Conference - IMC'2004*, Taormina, Sicily, Italy, octobre 2004.
- [KRI 00] KRISHNAMURTHY B., WANG J., « On network-aware clustering of web clients », *SIGCOMM*, 2000, p. 97-110.
- [MOO 00] MOORE D., PERIAKARUPPAN R., DONOHOE J., CLAFFY K., « Where in the World is netgeo.caida.org? », *Proc. of the INET'2000*, Yokohama, Japan, juillet 2000.
- [NLA] NLANR Active Measurement Project, <http://watt.nlanr.net/>.
- [PAD 01] PADMANABHAN V. N., SUBRAMANIAN L., « An Investigation of Geographic Mapping Techniques for Internet Hosts », *Proc. of the ACM SIGCOMM'2001*, San Diego, CA, USA, août 2001.
- [PER 03] PERCACCI R., VESPIGNANI A., « Scale-free behavior of the Internet global performance », *The European Physical Journal B - Condensed Matter*, vol. 32, n° 4, 2003, p. 411–414.
- [pla] « PlanetLab », <http://www.planet-lab.org>.
- [RIP] RIPE Test Traffic Measurements, <http://www.ripe.net/ttm/>.
- [Sar] Sarangworld Traceroute Project, <http://www.sarangworld.com/TRACEROUTE/>.
- [SUB 02] SUBRAMANIAN L., PADMANABHAN V. N., KATZ R., « Geographic Properties of Internet Routing », *Proc. of USENIX 2002*, Monterey, CA, USA, juin 2002.
- [Uni] University of Illinois at Urbana-Champaign, « IP Address to Latitude/Longitude », <http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll/>.
- [V.] V. Jacobson, Traceroute software, 1999, <ftp://ftp.ee.lbl.gov/traceroute.tar.Z>.
- [van 04] VAN LANGEN S., ZHOU X., VAN MIEGHEM P., « On the Estimation of Internet Distances Using Landmarks », *Proc. of the International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking – NEW2AN'04*, St. Petersburg, Russia, février 2004.
- [Vis] Visualware Inc., « VisualRoute », <http://www.visualware.com/visualroute/>.
- [YOO 02] YOOK S.-H., JEONG H., BARABÁSI A.-L., « Modeling the Internet's Large-Scale Topology », *Proc. of the National Academy of Sciences (PNAS)*, vol. 99, 2002, p. 13382–13386.
- [ZIV 04a] ZIVIANI A., FDIDA S., DE REZENDE J. F., DUARTE O. C. M. B., « Improving the Accuracy of Measurement-Based Geographic Location of Internet Hosts », *Computer Networks, Elsevier Science*, , 2004, Accepted for publication.
- [ZIV 04b] ZIVIANI A., FDIDA S., DE REZENDE J. F., DUARTE O. C. M. B., « Toward a Measurement-based Geographic Location Service », *Proc. of the Passive and Active Measurement Workshop - PAM'2004*, Lecture Notes in Computer Science (LNCS), Antibes Juan-les-Pins, France, avril 2004.