Arbres des plus courts chemins multiples dans l'Internet IPv6

Mickaël Hoerdt, Damien Magoni, Jean-Jacques Pansiot

Equipe Réseaux et Protocoles, LSIIT - Pôle API Boulevard Sébastien Brant 67400 ILLKIRCH FRANCE

Le protocole de routage multipoint le plus largement utilisé aujourd'hui est PIM-SM. Selon les cas, ce protocole permet la construction d'arbres partagés ou par source en se basant sur les tables de routage unicast IP. A l'heure actuelle, ces arbres sont uniques pour un groupe donné, car le protocole ne prend pas en compte l'existence de routes *unicast* de même coût (Equal Cost Multipath) vers une destination lors de leur construction. Or avec la prolifération du *multi-homing* et l'émergence d'IPv6, l'existence de plus courts chemins multiples dans l'Internet risque d'être de plus en plus forte. Dans cet article, nous proposons une modification de PIM-SM permettant de tirer parti des plus courts chemins multiples dans l'Internet. Nous justifions notre démarche par des mesures effectuées sur une topologie IPv6 réelle ayant les caractéristiques classiques des topologies Internet ainsi que par la simplicité de mise en oeuvre de notre proposition.

Keywords: multipoint, plus courts chemins multiples, extensibilité, déploiement, répartition de flux, IPv6

1 Introduction

Les propriétés des communications multipoints ont des implications plus complexes sur le réseau que les communications points à points. En particulier, prévoir la répartition des flux de données multipoints sur un réseau aussi étendu que l'Internet reste un problème plus difficile que pour les flux point à point. Ceci est du au fait qu'une communication de groupe nécessite l'établissement préalable d'un arbre de diffusion entre les différents membres du groupe. Celui-ci permet de transporter le flux de données multipoint entre les participants et nécessite donc de manière évidente plus de ressources qu'une simple communication point à point.

A l'heure actuelle, les arbres de diffusion multipoints construits sont de type *Reverse Shortest Path Tree*. L'ensemble des protocoles de routage multicast existant ne savent construire qu'un seul arbre des plus courts chemins pour un même ensemble $\{racine, feuilles\}$. Il est pourtant possible de considérer l'ensemble des RSPT si on fait l'hypothèse raisonnable que sur un graphe ayant la taille et les propriétés de l'Internet il peut exister plusieurs plus courts chemins entre deux noeuds. En effet, il est clair que si il existe plusieurs plus courts chemins entre deux noeuds, alors il peut exister plusieurs arbres des plus courts chemins pour un même groupe.

Le problème du calcul de plus courts chemins multiples a été largement étudié dans les communications point à point sur l'Internet [NZ01]. Il en est de même pour les arbres des plus courts chemins. L'idée d'utiliser les plus courts chemins pour améliorer le routage multicast n'est pas neuve [AHF03], mais actuellement l'impact des plus courts chemins multiples sur les arbres des plus courts chemins n'a pas encore été étudié dans la littérature. En particulier pour ce qui conçerne la quantification réaliste du nombre de plus courts chemins multiples et de leur usage possible lors d'une communication de groupe.

Plusieurs algorithmes distribués de construction de tables de routage point à point proposent déjà une gestion de plus courts chemins multiples comme par exemple IS-IS ou OSPF. Sur un routeur IP, cela se traduit par plusieurs prochains sauts vers une destination, ce qui peut être utilisé non seulement lors de la commutation de paquet unicast mais aussi lors de la construction d'un arbre de routage multipoint. Sur une topologie Internet IPv6 réelle établie lors de travaux précédents [HM04] nous cherchons à obtenir la répartition du nombre de plus courts chemins par feuilles dans un arbre donné pour des densités variées. En section 2 nous justifions l'utilité des plus courts chemins multiples dans un contexte de réseau à grande échelle. La section 3 présente des résultats de mesures effectuées pour vérifier nos hypothèses et la section

4 propose une modification simple du protocole de routage multicast le plus répandu permettant de tirer avantage des plus courts chemins multiples dans l'Internet. Nous concluons sur nos travaux en section 5.

2 Contexte

L'utilisation des plus courts chemins multiples lors de communications de groupes se justifie premièrement dans un contexte de déploiement à grande échelle car elle se base sur des mécanismes déjà existants et bien maitrisés : le calcul des plus courts chemins multiples en point à point et la commutation de paquet en saut par saut de type Internet. Deuxièmement elle se justifie dans le contexte des communications de groupe inter-domaine c'est à dire avec un nombre important d'arbres de diffusion dont la répartition des flux n'est a priori pas connue. La figure 1 illustre très simplement l'avantage de l'utilisation des plus courts chemins multiples de manière dynamique. Elle représente deux diffusions multipoints avec les mêmes membres mais des racines différentes. On voit bien que de manière évidente que le flux provenant de *m* aurait pu utiliser le même arbre de diffusion des plus courts chemins que le flux provenant de *k*. Mais un autre arbre est construit, afin de permettre une meilleure répartition de charge et un meilleur partage des ressources réseaux. Il faut bien voir qu'on se place dans un contexte de grande échelle, ce qui implique un nombre important d'arbres et de groupes possibles. Ainsi la simplicité du mécanisme proposé est à mettre en rapport avec le gain qu'on obtiendrait en le déployant sur un réseau du comme l'Internet.

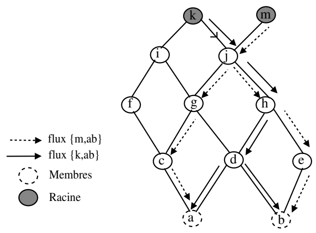


FIG. 1: Exemple de répartition de flux en utilisant les plus courts chemins multiples

Pour résumer, notre approche offre les avantages suivants :
- Incrémentale par rapport à l'existant et donc simple à déployer.
- Indépendante du type d'arbres construits.
- Indépendante du protocole de routage unicast sous-jacent.
- Permet une répartition des flux de manière dynamique.
- Permet un calcul de chemins de "secours" dans un arbre de diffusion.

3 Expérimentations

Les expérimentations effectuées par simulation sont sensibles à un nombre important de paramètres, en particulier celui de la topologie lorsqu'on cherche à trouver le gain d'une proposition à grande échelle. Dans cette section nous détaillons la méthodologie expérimentale ainsi que les caractéristiques importantes de la topologie utilisée.

3.1 Topologie

Afin de justifier notre approche, nous proposons dans cette section des mesures réalisée sur la topologie Internet IP version 6 obtenue lors de précédents travaux [HM04]. A notre connaissance nous sommes les premiers à avoir cartographié le réseau IPv6 et les seuls à utiliser cette topologie pour effectuer des expérimentations. Cette topologie nous garantit des résultats fiables car elle a été obtenue à l'aide de mesures réelles. La taille encore relativement petite de ce réseau nous permet d'effectuer des simulations centralisées dans un temps raisonnable. La table 3.1 présente quelques caractéristiques importantes de cette topologie IPv6. Nous avons vérifié dans nos précédents travaux [HM04] que celle-ci posséde certaines caractéristiques typiques du réseau IPv4 notamment en ce qui concerne les distributions des degrés, des arbres et du nombre de plus courts chemins distincts.

Source	Nombre de routeurs	Nombre de liens	Degre moyen des routeurs	Distance moyenne	Date de création
nec IPv6	4256	14326	6,73	7,92	Juin 2003

TAB. 1: Caractéristiques de la carte IPv6

3.2 Méthodologie

Nous aimerions quantifier l'existence de plus courts chemins multiples lors de la construction d'arbres de diffusion. Pour cela, nous avons sélectionné aléatoirement une part variable (de 1% à 40%) de noeuds membres d'un groupe multicast dans la topologie IPv6 ainsi qu'un noeud racine. Nous appelons *densité*, le pourcentage des noeuds de la carte étant membres d'un groupe. Une sélection aléatoire est justifiée car il est assez difficile de prévoir l'utilisation future de la diffusion de groupe à l'échelle de l'Internet. Celle-ci n'est encore que peu déployée à l'échelle inter-domaine et peu d'applications existent à l'heure actuelle. Cette densité de noeuds membres représente donc les feuilles d'un arbre de diffusion qu'on pourrait construire en utilisant les plus courts chemins depuis chaque feuille jusqu'à la racine. Ensuite, pour chaque noeud, nous calculons le nombre de plus courts chemins distincts (et disjoints) jusqu'à la racine.

S'assurer du réalisme et de la précision de la topologie utilisée est vital pour des simulations concernant le routage multi-chemin, car les résultats sont très dépendants de la topologie sous-jacente. C'est pourquoi nous utilisons la topologie décrite dans la section précédente pour effectuer nos simulations. Cette carte construite par des traces distribuées est très précise sur la quantité et le placement de ses liens.

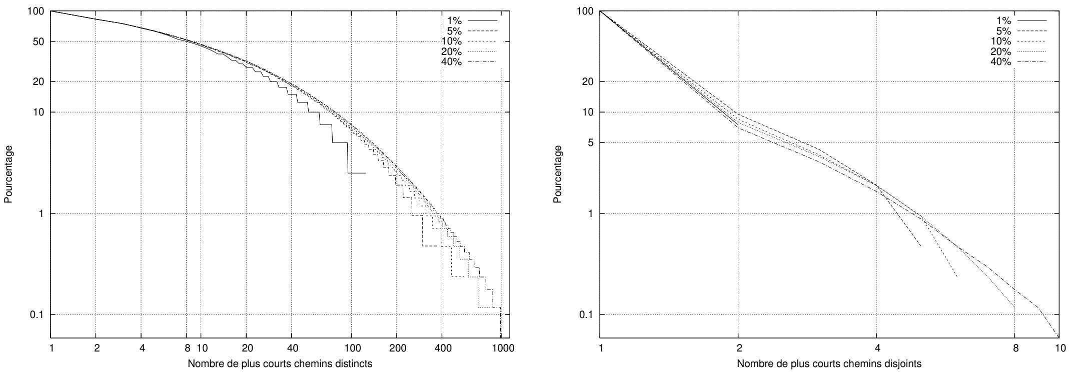
Comme le processus consistant à sélectionner des noeuds membres et racines fait intervenir des sélections aléatoires (et donc des tirages aléatoires), nous avons utilisé un scénario séquentiel de simulation [LK00] pour obtenir les résultats présentés dans la section suivante. Nous avons utilisé le générateur de nombres pseudo-aléatoires Mersenne Twister [MN98] pour produire les nombres aléatoires nécessaires à nos simulations. Comme les tirages aléatoires sont la seule source d'aléas dans les simulations, nous pouvons assumer que les données mesurées en sortie obéissent au théorème de la limite centrale (i.e. les valeurs des mesures d'une variable de sortie donnée forment une distribution normale). Nous avons effectué des simulations finies dans lesquelles chaque expérience consiste à sélectionner aléatoirement des membres selon une densité fixée en paramètre ainsi qu'une racine et à calculer le nombre de plus courts chemins distincts et disjoints de chacun des membres vers la racine. Nous avons ensuite classé tous les membres en fonction de leur nombre de plus courts chemins distincts (et disjoints dans une deuxième phase) puis nous avons répété l'expérience jusqu'à la convergence du nombre de plus courts chemins des membres classés. Le résultat d'une simulation est donc la répartition typique des membres selon leur nombre moyen de plus courts chemins distincts/disjoints pour une densité donnée. Tous les résultats des simulations ont été obtenus en utilisant un niveau de confiance de 0.95 avec une erreur statistique relative de 5% pour toutes les variables de sortie. Nous avons implémenté notre expérimentation en C++ avec le logiciel Network Manipulator.

3.3 Résultats

Les graphes 2(a) et 2(b) tracent sur une échelle logarithmique la répartition des noeuds membres du groupe en fonction de leur nombre de plus courts chemins distincts et disjoints vers la racine. La première chose qu'on peut remarquer, mais qui est normale est que les résultats ne varient pas ou très peu en fonction de la densité pour les densités élevées. Pour les densités faibles, les résultats varient en fonction de la densité mais il s'agit d'un biais statistique. En effet nos mesures se focalisent sur des statistiques concernant une population de plus courts chemins dans laquelle l'une des extrémités est fixe. Dans un contexte d'arbres de diffusion, cette extrémité fixe est la racine de l'arbre et ces plus courts chemins sont autant de candidats potentiels pour faire passer le flux multicast. Statistiquement, la proportion des paires ayant un nombre de plus courts chemins élevé est plus faible lorsque la densité est faible, et leur pourcentage n'est pas assez significatif pour être pris en compte par la méthode de convergence décrite en section 3.2.

Le graphe 2(a) trace le pourcentage de membres ayant un nombre de plus courts chemins **distincts** supérieur ou égal à une valeur donnée. De manière évidente 100% des membres ont au moins 1 plus court chemin **distinct** vers la racine. Un peu moins de 50% des membres ont plus de 10 plus courts chemins distincts vers la racine. Ce résultat est encourageant lorsqu'on voit qu'il reste encore 20% des membres ayant plus de 40 plus courts chemins distincts. Le nombre d'arbres possible reste donc toujours très élevé et nous encourage fortement à étudier un mécanisme permettant aux protocoles de routage multicast d'en bénéficier.

Sur le graphe 2(b), nous traçons le pourcentage de membres ayant un nombre de plus courts chemins **disjoints** supérieur ou égal à une valeur donnée. Les valeurs sont beaucoup plus faibles que sur le graphe



(a) Part du nombre de membres ayant un nombre de plus courts chemins **distincts** supérieur ou égal à une valeur donnée en fonction de la densité du groupe

(b) Part du nombre de membres ayant un nombre de plus courts chemins **disjoints** supérieur ou égal à une valeur donnée en fonction de la densité du groupe

FIG. 2: Répartition du nombre de chemins par membres

précédent et permettent de se rendre compte que même si il existe un nombre important de plus courts chemins distincts, il se peut que ceux-ci partagent un nombre important d'arêtes communes. Néanmoins, environ 10% des noeuds ont au moins 2 plus courts chemins disjoints jusqu'à la racine, ce qui n'est pas rien compte tenu du fait que tous les noeuds de l'arbre ont été tiré aléatoirement. En résumé nous pensons que le nombre de plus courts chemins mesurés par arbre sur une topologie réelle IPv6 est suffisamment intéressant pour imaginer une mécanisme simple en tirant partie.

4 Construction de routes alternatives

Nous proposons ici une modification du protocole PIM [FHHK03] permettant de calculer l'ensemble des plus courts chemins depuis chaque membre du groupe jusqu'à la racine. PIM pour Protocol Independent Multicast est un protocole de contruction d'arbre de diffusion multicast. Celui-ci utilise les tables de routage point à point pour construire de manière complètement distribuée et asynchrone un arbre de routage depuis les feuilles de l'arbre qui sont en général des machines hôtes, vers la racine qui est soit un hôte soit un point de rendez vous. Actuellement lorsqu'un noeud PIM détecte plusieurs prochains sauts vers la racine dans la table de routage point à point, il n'envoie qu'un seul message *join* vers l'un des prochains sauts en direction de la racine. Pour maintenir l'ensemble des plus courts chemins depuis chaque membre, il suffit simplement d'envoyer un message *join* pour chaque prochain saut indiqué dans la table de routage unicast comme illustré sur la figure 3. On n'obtient plus alors un arbre des plus courts chemins et, utilisé comme tel, il est clair que l'on s'expose à des risques de duplication de paquets. Pour éviter cela, il suffit de marquer les messages *join*. On distingue les messages *join(mandatory)* et *join(optional)*. Les messages *join(mandatory)* construisent effectivement l'arbre de diffusion sur lequel le flux multipoint sera transmis tandis que les messages *join(optional)* construisent des plus courts chemins de "secours". Un noeud recevant un message *join(mandatory)* sur l'une de ses interfaces le propagera en direction de la source de diffusion multipoint. Un noeud ne recevant que des messages *join(optional)* sur ses interfaces ne propagera que des messages *join(optional)* en direction de la source. Le mécanisme est illustré en figure 3. On voit donc que notre proposition permet de maintenir un arbre de diffusion PIM actif tout en "marquant" tous les plus courts chemins alternatifs uniquement en se basant sur les tables de routage point à point fournies par les protocoles de routage unicast sous-jacents. Ces chemins alternatifs peuvent être utilisés pour reconfigurer un arbre de diffusion à la volée, en fonction de la charge

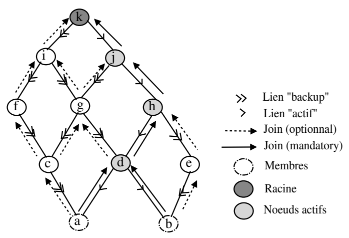


FIG. 3: Construction de l'ensemble des plus courts chemins pour un arbre PIM

présente sur tel ou tel lien. En effet si on reprend l'exemple de la figure 3 rien n'empêche le noeud *a* d'envoyer un message *join(optional)* sur le lien *ac* plutôt que *ad* si il se rend compte qu'il subit trop de pertes par ce même lien. De même pour le noeud d, si il se rend compte d'un début de congestion sur le lien *db*. On peut se demander l'utilité de construire des chemins de "secours" en même temps que l'établissement de l'arbre de diffusion et pas au moment même ou l'on détecte une anomalie dans le réseau. Nous construisons les chemins avant afin de traiter non seulement les cas d'anomalies, mais aussi les basculements d'un arbre vers un autre ayant des performances plus intéressantes.

Une critique de notre approche pourrait être que la maintenance des chemins de "secours" est trop coûteuse par rapport au gain optenu mais nous avons montré d'une part dans des travaux précédents [HPMG03] que certains noeuds de l'Internet risquaient fortement de supporter une part importante du trafic multicast IP à l'avenir. D'autre part, le nombre de messages *join(optional)* envoyé en direction de la racine peut être décidé localement et n'a pas forcément l'obligation d'être égal au nombre de prochains sauts indiqués dans la table de routage unicast.

5 Conclusion et perspectives

Dans cet article, nous avons analysé l'impact de l'utilisation des plus courts chemins multiples dans les communications de groupes. En particulier nous nous sommes focalisés sur les mesures concernant un groupe (et donc un arbre) en fonction de sa densité. Ces mesures ont été effectuées sur une topologie réelle, celle de l'Internet IPv6 actuel. Dans une optique de facilité de déploiement, nous avons montré que l'utilisation des plus courts chemins multiples dans un arbre de diffusion multipoint est réellement envisageable. En effet nous avons montré qu'il est possible de modifier le protocole de routage multipoint le plus utilisé actuellement pour calculer ces plus courts chemins multiples reliants les membres d'un groupe à la racine de leur arbre de diffusion. Cette modification offre un réel gain pour la diffusion de groupe à l'échelle de l'Internet inter-domaine au vu de sa simplicité et des résultats de nos simulations. Nous espérons poursuivre nos expérimentations, notamment sur la reconfiguration automatique d'un arbre de diffusion multipoint en fonction de la charge de trafic subie par un noeud. Après avoir identifié clairement les critère de reconfiguration, nous pensons pouvoir implémenter la solution très rapidement.

Références

[AHF03] José Ignacio Alvarez-Hamelin and Pierre Fraigniaud. A multicast protocol with qos support. In *Proceedings of the 12th Conference on Computer Communication and Networks (IC3N)*, Dallas,Texas, oct 2003.

[FHHK03] Bill Fenner, Mark Handley, Hugh Holbrook, and Isidor Kouvelas. "draft-ietf-pim-sm-v2-new-08.txt" : Protocol independent multicast-sparse mode (PIM-SM) : Protocol specification (revised). http ://www.ietf.org/internet-drafts/draft-ietf-pim-dm-new-v2-08.txt, October 2003.

[HM04] Mickaël Hoerdt and Damien Magoni. Distribution of multicast tree states over the ipv6 network topology. In *Proceedings of the IEEE International Conference on Communications*, Paris, France, June 2004. To appear.

[HPMG03] Mickaël Hoerdt, Jean-Jacques Pansiot, Damien Magoni, and Dominique Grad. Répartition des états de routage multipoint par les plus courts chemins multiples. In *Actes des 5èmes rencontres francophones sur les aspects Algorithmiques des Télécommunications*, pages 139–145, Banyuls-sur-mer, France, Mai 2003.

[LK00] A.M. Law and W.D. Kelton. *Simulation Modelling and Analysis*. McGraw-Hill, 3rd edition, 2000.

[MN98] Makoto Matsumoto and Takuji Nishimura. Mersenne twister : a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1) :3–30, 1998.

[NZ01] Srihari Nelakuditi and Zhi-Li Zhang. On selection of paths for multipath routing. In *IWQoS*, Karlsruhe, Germany, June 2001.