ANALYSE DES SOLUTIONS EXISTANTES POUR UN ROUTAGE PAIR-À-PAIR EFFICACE

Ce chapitre analyse les différentes solutions existantes à base de DHT pour la prise en compte du réseau sous-jacent.

Le terme 'proximité' désigne une distance proche dans le réseau sous-jacent, en termes de sauts IP.

3.1 - Techniques d'exploitation de la proximité

Les premières DHTs ont été développées indépendamment de la topologie sous-jacente. Toutefois, Pastry [RD01] utilise certaines heuristiques pour explorer la proximité des pairs dans le réseau sous-jacent afin de pouvoir construire ses tables de routage *overlay* (notamment le *neighborhood set*). Cependant les impacts sur le routage P2P sont assez peu importants puisqu'une décision de router vers un pair de cette table de proximité n'est prise qu'en dernier recours face à un choix de nœuds. La localité dans Pastry n'est donc pas exploitée pour une stratégie de prise en compte du réseau sous-jacent.

Les techniques existantes qui permettent d'exploiter les informations de proximité dans le réseau sous-jacent sont classées en trois catégories [CDH+02] :

- le routage de proximité (*proximity routing*) ;
- l'agencement géographique, qui sous-entend une affectation du *nodeId* en fonction de la topologie (*topology-based nodeId assignment*) ;
- et le choix du voisin à proximité (*proximity neighbor selection* ou *PNS*) [RSS02].

3.1.1 - Routage de proximité

Le routage de proximité s'applique à un réseau *overlay* construit indépendamment du réseau sous-jacent. Durant le mécanisme de routage, si un pair a l'opportunité de choisir parmi *k* prochains sauts possibles, il choisit de router le message vers le nœud qui, parmi ces *k* nœuds, est le plus proche de lui dans le réseau sous-jacent. Une alternative serait de choisir de router vers le nœud qui représente le meilleur compromis entre la proximité et la progression dans l'espace d'identification *overlay*. En tout cas, les performances globales de cette technique dépendent de *k*, qui est proportionnel à la taille de la table de routage. Par ailleurs, des petits sauts risquent d'être contrecarrés par l'augmentation du nombre de sauts.

3.1.2 - Agencement géographique

La technique de l'agencement géographique vise à reporter l'espace d'identification *overlay* sur la topologie sous-jacente, et biaise l'attribution des *nodeId*s. Ceci entraîne forcément une diminution des délais, mais les *nodeId*s ne sont plus uniformément distribués dans l'espace d'identification *overlay*. D'où des problèmes d'équilibrage de charge. Les nœuds voisins sont aussi susceptibles de subir des pannes ou attaques corrélées, ce qui affaiblit la robustesse et la sécurité du système. Malgré ces revers, cette technique a bien permis de créer une version de CAN [RFH$^+$01, Rat02] sensible à la topologie [RHK$^+$02]. Elle ne peut toutefois pas être appliquée à des espaces d'identification à une seule dimension, comme est le cas de la majorité des DHT (e.g. Chord [SML$^+$01], Pastry [RD01], etc.)

3.1.3 - Choix du voisin à proximité

Comme la technique d'agencement géographique, la technique du choix du voisin à proximité (*proximity neighbor selection* ou *PNS*) [RSS02] intervient dans la construction d'un *overlay* conscient

de la topologie sous-jacente. Mais au lieu de biaiser l'attribution des *nodeId*s, cette technique choisit pour les entrées de sa table de routage des références vers les nœuds de son voisinage qui satisfont les contraintes algorithmiques du protocole mis en œuvre. La technique *PNS* n'est alors susceptible de satisfaire ses ambitions que si elle est appliquée à un algorithme qui permet une certaine liberté dans la construction des tables de routage sans affecter le diamètre du réseau. Elle peut donc être appliquée à la construction du *neighborhood set* de Pastry [RD01], par exemple, mais pas à la construction de DHTs CAN [RFH+01, Rat02] ou Chord [SML+01], où chaque entrée de la table de routage fait référence à un point bien précis de l'espace d'identification *overlay*. Mais quand elle est mise en œuvre, la technique *PNS* introduit un surcoût assez faible et facilite la gestion de l'antémémoire (puisque les nœuds voisins dans le réseau *overlay* sont alors susceptibles d'être voisins au niveau sous-jacent). Cette technique de choix du voisin à proximité s'adapte à des caractéristiques de réseau variables. Elle peut être considérée aussi comme un bon compromis qui diminue le diamètre du réseau et préserve l'équilibre de charge. Cependant, la découverte des voisins est étroitement liée au protocole.

3.2 - Solutions de prise en compte de la topologie du réseau sous-jacent

Les trois techniques d'exploitation de l'information de proximité présentées au paragraphe précédent ont des limites importantes, et parfois s'impliquent dans le protocole de routage *overlay*. Voilà pourquoi de nouvelles solutions structurelles ont émergé. Elles sont nombreuses et la liste continue à s'allonger. Dans ce paragraphe, nous présentons brièvement les principaux systèmes P2P utilisant des DHTs et connus pour prendre en compte la topologie du réseau sous-jacent, à savoir : Brocade [JZD+02], l'*expressway* (la voie expresse) [XMK03], Hieras [XMH03], Toplus [GER+03] et Plethora [FGJ04].

3.2.1 - Brocade

Brocade [JZD+02] construit un réseau *overlay* secondaire au-dessus d'un réseau P2P utilisant une DHT. Ce réseau secondaire exploite les caractéristiques du réseau sous-jacent au réseau P2P initial. Il est constitué en fait des pairs ayant de grandes capacités réseau (en termes de bande passante, puissance du processeur, etc.) et situés près des points d'accès au réseau IP (e.g. passerelles, routeurs). Ces pairs particuliers sont dits des *super-nœuds* et agissent comme des points de repère (*landmarks*) pour chaque domaine du réseau IP. Chaque *super-nœud* gère un groupe de pairs locaux afin de réduire le trafic dans le réseau. En même temps, chaque *super-nœud* garde des informations de tous les pairs de son domaine ; ce qui peut les engorger.

Pour router un message dans Brocade, un pair se connecte au *super-nœud* le plus proche et utilise le réseau *overlay* secondaire comme raccourci vers la destination. L'usage de la bande passante est ainsi réduit, de même que le nombre total de sauts IP. Sauf que le réseau secondaire de Brocade utilise toujours un routage logique qui ne prend pas en compte la topologie sous-jacente. Du coup, Brocade ne résout le problème qu'à un certain degré et le ramène à un réseau secondaire plus petit (en diamètre et degré).

3.2.2 - Expressway

L'*expressway* (la voie expresse) [XMK03] est un réseau auxiliaire à un réseau P2P existant utilisant les DHTs. Il est construit au-dessus de ce réseau dans le but d'accélérer le routage *overlay* en tirant profit de l'hétérogénéité inhérente au réseau IP sous-jacent. Cette hétérogénéité est assez diversifiée : en termes de connectivité des nœuds, leur proximité physique relative, leur disponibilité, leur capacité de transmission, etc.

L'*expressway* est formée par des *expressway nodes*. Ce sont des pairs de grandes capacités et situés près des passerelles ou près

des routeurs. Pour la construction de cette voie expresse, il existe deux approches génériques.

- La première approche utilise la topologie des niveaux des ASs, dérivée des tables BGP [RFC1771]. Pour router un message, un pair commence par contacter l'*expressway node* qui se trouve dans son AS. Cette approche nécessite de chaque pair de l'*expressway* de connaître tous les nœuds de son AS. De plus, les revers de cette approche sont similaires à ceux de Brocade [JZD+02].
- La deuxième approche utilise une technique de numérotage de points de repère par groupement de nœuds et exploite la proximité sous-jacente par la technique *PNS* [RSS02], discutée ci-dessus (paragraphe 3.1.3). Pour router un message, un pair commence par contacter l'*expressway node* qui se trouve dans son groupement. Le routage se poursuit par une technique de distance vectorielle.

3.2.3 - Hieras

Hieras [XMH03] est un système hiérarchique multi-niveau destiné à remédier à un routage *overlay* indifférent aux temps de latence dans le réseau sous-jacent. Chaque pair appartient simultanément à tous les niveaux et gère autant de listes ou tables nécessaires à son appartenance à chaque niveau.

Au niveau supérieur de l'architecture, tous les pairs actifs sont regroupés dans un même ensemble, appelé *ring*. C'est le niveau du réseau P2P existant ; les autres niveaux s'intercalent donc au-dessus du réseau *underlay* initial. À chacun de ces niveaux *overlay*s intermédiaires, les pairs adjacents (au sens topologique) sont groupés en plusieurs *ring*s disjoints. Chacun de ces *ring*s constitue un sous-groupe du réseau P2P global. L'algorithme de routage dans chaque *ring* est le même que celui mis en œuvre au niveau global.

Les différents *ring*s des niveaux intermédiaires sont construits de sorte que le temps de latence moyen entre deux quelconques de leurs pairs décroît en passant d'un niveau intermédiaire à un autre qui lui est inférieur. Ce temps de latence sera donc minimal au niveau *overlay* le plus bas, et les *ring*s y sont le plus petit.

Afin de pouvoir estimer une telle proximité pour chaque *ring*, Hieras emploie une technique de bacs distribués (ou *distributed binning*) [RHK+02] qui nécessite l'existence de nœuds de marquage bien définis.

Le mécanisme de routage s'exécute en premier au niveau le plus bas, dans le *ring* de l'initiateur de la requête. En cas d'échec, le routage reprend dans le *ring* du niveau supérieur, et ainsi de suite jusqu'à satisfaction de la requête. Généralement, la majorité des requêtes sont satisfaites à un niveau intermédiaire, réduisant ainsi les temps de latence puisque les requêtes n'atteignent donc plus le réseau global au niveau le plus haut.

3.2.4 - Toplus

Toplus [GER+03] est un service hiérarchique de recherche de données dans les réseaux P2P utilisant des DHTs. Il regroupe les pairs selon le préfixe IP de leur réseau et organise les groupes hiérarchiquement en de nouveaux groupes suivant la topologie hiérarchique de l'Internet récupérée des tables BGP [RFC1771]. Ainsi un AS est-il subdivisé en une hiérarchie IP.

Le mécanisme de routage de Toplus est basé sur une généralisation de la règle du plus long préfixe correspondant. Comme Kademlia [MM02] (cf. p. 54), il utilise la métrique du choix exclusif, XOR.

La structure et les performances de Toplus suivent bien celles du réseau Internet. Cependant, ce système présente un certain nombre d'inconvénients. Évidemment, il est susceptible de subir des pannes de nœuds corrélées qui peuvent faire tomber tout un groupe

de nœuds dont les adresses IP sont reliées. Par ailleurs, le nombre de pairs actifs dans un (sous-)groupe n'est pas nécessairement proportionnel au nombre d'adresses IP qu'il couvre. La population de l'espace d'identification n'est donc pas uniforme, ce qui conduit à un déséquilibre de charge entre les pairs.

3.2.5 - Plethora

Plethora [FGJ04] est un référentiel de données à large échelle, structuré en deux niveaux. Au niveau inférieur, le réseau *overlay* global contient tous les pairs. Au-dessus, le cœur de routage de Plethora organise les pairs en plusieurs *overlay*s locaux selon leur proximité relative et en conformité avec les AS. Les *overlay*s locaux sont tous à un même niveau et servent d'antémémoires de la localité pour l'*overlay* global afin d'améliorer les temps d'accès aux éléments de données. Les requêtes sont alors routées en premier dans l'*overlay* local. Cependant, la taille des *overlay*s locaux impacte les performances du système. Pour cela, elle est définie par des paramètres de système, constamment contrôlée et régulée au besoin par deux algorithmes distribués :

- les paramètres systèmes fixent un minimum et un maximum de nombre de pairs par *overlay* local ;
- dans chaque *overlay* local, le pair à l'identifiant le plus petit se charge de contrôler sa taille ;
- au besoin, un algorithme se charge d'unifier deux *overlay*s locaux devenus petits ;
- et si besoin y est, un autre algorithme se charge de partager un *overlay* local en garantissant avec une grande probabilité que les pairs issus d'un même AS restent dans un même *overlay* local.

Plethora suppose des pairs ayant une bonne connectivité Internet en termes de bande passante, et étant partiellement statiques.

3.3 - Solutions de coopération entre flux pair-à-pair et opérateur de réseaux

De nombreux systèmes ont émergé visant à améliorer les performances d'un système P2P utilisant une DHT, en exploitant l'information de proximité sous-jacente. Deux architectures proposent (indépendamment des DHTs) une coopération inter-couches entre le trafic P2P et les politiques d'un opérateur de réseaux. L'opérateur de réseau considéré dans ces deux propositions est un ISP.

3.3.1 - Solution de service intermédiaire

Une interaction active entre le réseau P2P et l'infrastructure sous-jacente d'un ISP est possible en passant par le service d'un serveur intermédiaire [AFS07]. Ce serveur, appelé *oracle*, est hébergé par l'ISP et aident les usagers P2P à choisir des pairs de façon optimale, notamment dans le cas d'une application P2P de téléchargement de fichiers. Étant géré par l'ISP, l'*oracle* a un accès direct à toutes les informations nécessaires (e.g. celles relative à la topologie physique réelle).

Avant d'entrer en lien avec un autre pair, chaque pair envoie à l'*oracle* une liste des pairs potentiellement voisins. L'*oracle* ordonne alors la liste selon un certain nombre de critères que chaque ISP est libre de spécifier. Parmi ces critères, l'on peut citer : la proximité ou le taux de bande passante disponible par lien (entre un pair de la liste et le pair qui a construit la liste), les politiques de routage, les termes des accords signés avec d'autres ISPs, etc.

L'*oracle* agit donc à la fois comme une couche de routage abstraite sous-jacente au réseau P2P et comme un service offert par l'ISP. Il peut être implémenté de façon distribuée afin de pallier les risques d'engorgement.

3.3.2 - Provider Portal for Peer-to-Peer

P4P (*Provider Portal for P2P*) [XKS+07] est une architecture légère permettant une communication explicite entre le trafic P2P et

les ISPs dans le cadre des applications à partage de fichiers. Elle vise à réduire le trafic dans la dorsale (*backbone*) et diminuer les coûts d'opérations. Elle s'appuie sur le fait qu'un ISP est le mieux placé pour déterminer la localité et router les messages non seulement vers des pairs voisins mais aussi vers des pairs accessibles par des liens à meilleur débit et peu chargés.

La structure de P4P consiste en un plan de données et un plan de contrôle avec une sorte d'entité centrale appelée *iTracker*. Cet *iTracker* fournit trois types d'informations concernant l'ISP :

– la topologie et le statut du réseau ;
– les politiques et directives de l'ISP ;
– et les capacités du réseau.

Avec son *iTracker*, P4P semble être une sorte d'architecture centralisée appliquant une architecture CDN (cf. paragraphe 1.2.4.2 p. 29) à un réseau P2P de partage de fichiers.

CONCLUSION DU CHAPITRE –

Ce chapitre a analysé les techniques existantes d'exploitation de l'information de proximité sous-jacente, ainsi que les principales solutions de prise en compte de la topologie sous-jacente dans les systèmes P2P à base de DHT. Ces techniques et solutions ne permettent cependant pas une coopération entre le flux P2P et les opérateurs de réseaux. Le chapitre se poursuit donc par l'analyse de deux solutions émergentes expressément dédiées à une telle coopération. Or il s'avère que l'une nécessite une étude systématique de chaque requête, et l'autre nécessite une infrastructure particulière et ne respecte pas la parité pure recherchée entre les nœuds du système pour un routage P2P efficace.

Les chapitres suivants sont dédiés à nos contributions en vue d'une sensibilisation efficace d'un réseau P2P utilisant une DHT à l'infrastructure du réseau sous-jacent, avec une coopération possible entre les flux P2P et les opérateurs de réseaux.