RFC 6908 : Deployment Considerations for Dual-Stack Lite

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 26 mars 2013

Date de publication du RFC : Mars 2013

https://www.bortzmeyer.org/6908.html

DS-Lite est une des nombreuses techniques de transition / co-existence IPv4 / IPv6. Elle est normalisée dans le RFC 6333 [1]. Ce nouveau RFC se penche plutôt sur des problèmes opérationnels : comment déployer DS-Lite, quels pièges éviter, etc.

DS-Lite ("*Dual-Stack Lite*") est conçu pour un FAI récent, qui vient d'arriver et n'a donc pas pu obtenir d'adresses IPv4, toutes épuisées <https://www.bortzmeyer.org/epuisement-adresses-ipv4.html>. Ce FAI a donc dû numéroter son réseau en IPv6, et comme ses clients ont encore de l'IPv4 à la maison, et comme ils souhaitent accéder à des machines Internet qui sont encore uniquement en IPv4, le FAI en question va donc devoir 1) NATer les adresses de ses clients (RFC 3022) vers le monde extérieur 2) tunneler leurs paquets depuis leur réseau local IPv4 (RFC 2463) jusqu'au CGN, en passant par le réseau purement IPv6 du FAI. C'est là que DS-Lite intervient (RFC 6333).

Apparemment, il n'y a eu qu'un seul déploiement effectif de DS-Lite chez un vrai FAI (le RFC ne dit pas lequel) et ce RFC 6908 est largement tiré de cette expérience réelle. Il est donc une lecture intéressante pour les opérateurs qui aiment le concret.

Ce RFC est en deux parties : les questions liées à l'AFTR et celles liées au B4. Vous ne savez pas ce qu'est un AFTR ou un B4 ? Vous n'avez pas envie de lire le RFC 6333 qui explique cela ? Alors, en deux mots, l'AFTR ("*Address Family Transition Router*") est le gros routeur NAT (le CGN) situé entre le FAI et l'Internet. Et le B4 ("*Basic Bridging Broadband element*") est chez le client, typiquement dans sa box et représente le point d'entrée du tunnel, du « lien virtuel » ("*soft wire*"). Prononcez les sigles à voix haute : le B4 est avant le tunnel ("*before*") et l'AFTR après ("*after*").

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc6333.txt

Donc, bien qu'il soit « après », notre RFC commence par l'AFTR (section 2). D'abord, comme DS-Lite utilise un tunnel, il va y avoir des problèmes dus à la MTU inférieure. Le RFC recommande donc d'essayer de faire un lien entre le B4 et l'AFTR ayant une MTU d'au moins 1540 octets pour que, après avoir retiré les 40 octets de l'encapsulation dans le tunnel, on ait les 1500 octets d'Ethernet. Si ce n'est pas possible, alors le B4 et l'AFTR vont devoir gérer la fragmentation. Comme le réassemblage des paquets est une opération coûteuse en ressources (il faut garder les fragments en mémoire tant que tout n'est pas arrivé), la tâche va être particulièrement lourde pour l'AFTR (des milliers de tunnels peuvent se terminer sur un seul AFTR). Il faut les dimensionner largement ! (C'est un problème général des CGN.)

Ensuite, il faut penser à la journalisation. Le principe de DS-Lite est que tous les clients du FAI partageront la poignée d'adresses IPv4 publiques que NATeront les AFTR. Si un client fait quelque chose de mal (par exemple accéder à un site Web qui déplait au gouvernement), il faut pouvoir l'identifier. Un observateur situé à l'extérieur a vu une adresses IPv4 du FAI et des milliers de clients pouvaient être derrière. L'AFTR doit donc noter au moins l'adresse IPv6 du B4 (elle est spécifique à chaque client) et le port source utilisé après le NAT (en espérant que l'observateur extérieur ne notera pas que l'adresse IPv4 source mais aussi le port, cf. RFC 6302).

Les adresses IP partagées sont une plaie du NAT, ce n'est pas nouveau (RFC 6269). Par exemple, si une adresse IPv4 est mise sur une liste noire suite à son comportement <http://lci.tf1.fr/high-tech/2007-07/polemique-quand-wikipedia-punit-sciences-4890331.html>, ce seront des centaines ou des milliers de B4 qui seront privés de connectivité IPv4. Si le destinataire ne tient pas compte du RFC 6302 et bloque sur la seule base de l'adresse IPv4, le support téléphonique du FAI qui a déployé DS-Lite va exploser. (Un travail est en cours à l'IETF pour aider à l'identification de l'utilisateur derrière un CGN. Mais il n'est pas terminé et, de toute façon ne sera pas forcément déployé. L'obstination de tant d'acteurs à rester en IPv4 va se payer cher.)

Un problème analogue se posera si le FAI veut faire de la comptabilité des octets échangés, et s'il ne met pas cette fonction dans le CPE. Le tunnel et le NAT rendront cette analyse plus compliquée (section 2.6). À noter que le RFC 6519 normalise des extensions à Radius pour les problèmes particuliers de DS-Lite.

Mais tout ceci n'est rien par rapport aux problèmes de fiabilité que posent les AFTR. L'Internet normal est très robuste car les équipements intermédiaires (les routeurs) n'ont pas d'état : s'ils redémarrent, rien n'est perdu. S'ils s'arrêtent de fonctionner, on en utilise d'autres. Le NAT met fin à tout cela : le routeur NAT a un état en mémoire, la table des correspondances {adresse interne, adresse externe} et, s'il redémarre, tout est fichu. C'est bien pire pour un CGN : s'il redémarre, ce sont des centaines ou des milliers d'utilisateurs qui perdent leurs connexions en cours. La fiabilité des AFTR est donc cruciale (cf. annexe A.3 du RFC 6333). Notre RFC propose que des AFTR de remplacement soient installés pour être prêts à prendre le relais. Ils peuvent fonctionner en « remplacement à chaud » (”*Hot Standby*”), se synchronisant en permanence avec l'AFTR primaire, prêts à le remplacer « sans couture » puisqu'ils ont le même état, ou bien en « remplacement à froid » (”*Cold Standby*”), sans synchronisation. Dans ce second cas, l'AFTR de secours fait qu'un AFTR en panne est vite remplacé, mais il n'y a pas de miracle : les clients doivent recommencer toutes leurs sessions. Vue la complexité qu'il y a à mettre en œuvre correctement la synchronisation des états, notre RFC recommande le remplacement à froid, et tant pis si c'est moins agréable pour les utilisateurs.

Et les performances ? Combien faut-il d'AFTR pour servir les utilisateurs ? La section 2.8 discute de ce point et du meilleur placement de ces équipements.

Comme avec tout CGN, DS-Lite, en centralisant les adresses IPv4, limite les possibilités de géo-localisation. Le B4 (donc la maison de l'utilisateur) et l'AFTR peuvent parfaitement être dans des villes différentes (RFC 6269, section 7). Les applications ne devraient donc pas se fier à la géo-localisation mais

demander à l'utilisateur où il se trouve, par exemple avec le protocole HELD du RFC 5985 ou en suivant l'architecture du RFC 6280. Le RFC conseille, curieusement, de mettre les AFTR le plus près possible des utilisateurs, pour limiter ce problème (alors que les AFTR ont besoin d'une connectivité IPv4, et que le réseau du FAI ne la fournit pas forcément, sinon il n'aurait pas besoin de DS-Lite).

Autre problème douloureux avec les CGN (dont DS-Lite), les connexions entrantes. Contrairement au NAT traditionnel, celui fait par la "*box*" à la maison, plus question de configurer sa "*box*" via une interface Web pour rediriger certains ports. Et plus moyen de se servir d'UPnP. Les éventuelles redirections doivent être sur l'AFTR, un engin largement partagé. Donc, pas question de demander une redirection du port 80...

Pour « programmer » l'AFTR, lui demander d'ouvrir certains ports en entrée, notre RFC recommande d'installer un serveur PCP ("*Port Control Protocol*", RFC 6887) sur les AFTR. Il n'est pas évident que cela soit une solution satisfaisante. D'une part, il existe encore très peu de clients PCP (par rapport aux clients UPnP). D'autre part, l'AFTR étant partagé entre plusieurs clients ne se connaissant pas, rien ne dit que la coexistence se passera bien et que les opérateurs accepteront d'ouvrir PCP à leurs clients.

Voilà, c'étaient les points essentiels pour les AFTR, lisez la section 2 du RFC si vous voulez la liste complète. Et les B4 ? Ils font l'objet de la section 3. D'abord, le problème du DNS. Le B4 est censé s'annoncer comme résolveur DNS et relayer les requêtes DNS au dessus d'IPv6 vers l'extérieur. Ce faisant, il est important qu'il suive les recommendations pour les relais DNS du RFC 5625 (en pratique, la grande majorité des "*boxes*" et des petits routeurs bas de gamme violent affreusement ces recommendations). C'est notamment nécessaire si on veut que DNSSEC fonctionne.

Si un client sur un réseau local IPv4 fait des requêtes DNS en IPv4 vers une adresse qui n'est pas celle du B4, ces requêtes seront tunnelisées et NATées comme n'importe quel trafic IPv4. Compte-tenu du fait que, du point de vue du NAT, chaque requête DNS est une session à elle seule, l'augmentation de la charge sur les AFTR sera sensible. Ce n'est donc pas conseillé.

Le B4 étant le début du réseau de l'opérateur, et beaucoup de choses dépendant de sa configuration et de son bon fonctionnement, il est important que l'opérateur puisse y accéder à distance, et au dessus d'IPv6, pas au dessus du lien virtuel (qui a davantage de chances d'avoir des problèmes).

À propos d'administration et de supervision du réseau, il est important que les AFTR répondent aux paquets utilisés pour ping et traceroute, afin que le B4 puisse tester que la connectivité IPv4 au dessus du tunnel marche bien. Des adresses IPv4 spéciales ont été réservées pour cela (rappelez-vous que, si vous déployez DS-Lite, c'est que vous n'avez pas assez d'adresses IPv4 publiques), `192.0.0.0/29`. Les AFTR sont censés tous répondre aux paquets envoyés à `192.0.0.2`.