RFC 7586 : Scaling the Address Resolution Protocol for Large Data Centers (SARP)

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 28 juin 2015

Date de publication du RFC : Juin 2015

https://www.bortzmeyer.org/7586.html

Le problème de passage à l'échelle de protocoles de recherche d'adresse MAC des voisins, les protocoles comme ARP, sont connus depuis un certain temps, et documentés dans le RFC 6820 [1]. Résumé en deux mots, dans un grand centre de données non partitionné en sous-réseaux IP, le trafic ARP peut représenter une partie significative du travail à effectuer par les machines. Ce nouveau RFC expose une des solutions pour faire face à ce problème : SARP ("*Scaling the Address Resolution Protocol*") fait appel à des relais ARP qui peuvent générer localement la plupart des réponses.

Si le centre de données est rigoureusement découpé en sous-réseaux IP (par exemple un sous-réseau, et donc un routeur par baie), il n'y a pas de problème ARP : le trafic ARP reste local. Mais si on veut profiter de la souplesse que permet la virtualisation, par exemple en déplaçant des machines virtuelles d'un bout à l'autre du centre de données en gardant leur adresse IP, on doit alors propager les requêtes ARP sur une bien plus grande distance et les problèmes de passage à l'échelle apparaissent (RFC 6820). La mémoire consommée par la FDB ("*Filtering Data Base*", la table des adresses MAC connues) augmente, ainsi que le temps de traitement de tous ces paquets ARP diffusés.

Les premières versions des brouillons ayant mené à ce RFC ne mentionnaient qu'ARP (RFC 826), protocole de résolution IP-¿MAC pour IPv4. Mais la version finale considère que le protocole marche aussi bien pour ND (RFC 4861), son équivalent pour IPv6. Seul le nom de la solution garde trace de cette préférence pour ARP. Dans le reste de cet article, je parlerais de ARP/ND.

L'idée de base de SARP est que chaque domaine d'accès (un groupe de machines proches, par exemple dans la même baie ou dans la même rangée) ait un relais ("*SARP proxy*") qui connaisse les

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc6820.txt

adresses MAC de tout le domaine, et réponde aux requêtes ARP/ND pour les autres domaines avec sa propre adresse MAC. Ainsi, la taille de la table ARP des machines du domaine reste proportionnelle à la taille du domaine d'accès, pas au nombre total de machines (comme ce serait le cas avec un réseau « plat » classique, entièrement en couche 2, et sans SARP).

Le relais SARP peut être l'hyperviseur d'un groupe de machines virtuelles (commutateur virtuel) ou bien il peut être dans un commutateur physique, ToR ("*Top of Rack*") ou bien EoR ("*End of Row*"). En gros, le relais SARP est là où un domaine d'accès se connecte au cœur du réseau interne du centre de données. Ce doit être une grosse machine car elle va devoir stocker les adresses MAC de toutes les machines qui communiquent avec une machine d'un autre domaine d'accès. Et il peut aussi faire l'objet d'attaques délibérées (cf. section 4).

La section 3 de notre RFC décrit plus en détail le fonctionnement de SARP. Si la machine source et la destination sont dans le même domaine d'accès (même baie, ou même rangée, selon l'endroit où se trouve le commutateur), ARP/ND fonctionne comme d'habitude et SARP n'intervient pas. Si la machine de destination est dans un autre sous-réseau IP, on passe alors par le routeur, selon le mécanisme normal de la couche 3. Mais si la destination est dans le même sous-réseau IP, mais dans un domaine d'accès différent ? Le relais SARP voit alors passer la requête ARP/ND. Si la réponse est dans son cache (qui associe des adresses IP à des adresses MAC), il répond avec sa propre adresse MAC (ainsi, les machines du domaine d'accès local ne sont pas noyées par des milliers d'adresses MAC de tout le centre de données). Sinon, il transmet à tous les domaines d'accès qui peuvent avoir cette adresse IP puis relaie la réponse. Seuls les relais SARP ont un cache qui contient des adresses MAC de tout le centre de données. Les machines ordinaires n'ont que les adresses MAC de leur propre domaine d'accès.

Et pour transmettre un paquet de données ? La machine source, ayant reçu l'adresse MAC du relais SARP en réponse à sa requête ARP/ND va donc mettre sur le câble un paquet ayant pour adresse Ethernet de destination le relais SARP. Le relais SARP, utilisant son propre cache (qui, lui, est complet), remplace l'adresse MAC de destination par la « vraie », et l'adresse MAC source par la sienne (pour qu'une réponse puisse revenir), et remet le paquet sur le câble.

Un tel mécanisme fait que des opérations comme la migration d'une VM d'un bout à l'autre du centre de données sont complètement invisibles. Les mécanismes normaux de résolution feront tout le travail. Cela suppose toutefois que la machine qui se déplace (ou plutôt son hyperviseur qui, contrairement à la VM, est conscient du déplacement) émette tout de suite un paquet ARP gratuit ou un paquet ND non sollicité, pour que les caches soient mis à jour (autrement, la machine migrée restera injoignable le temps que l'entrée dans le cache expire).

Une conséquence de cette technique est que le relais SARP est absolument vital : s'il est en panne, plus rien ne marche, à part les communications locales à un domaine d'accès. Il vaut donc mieux en avoir plusieurs, pour chaque domaine d'accès.

Ce RFC n'a que le statut « Expérimental » car l'IESG n'est pas convaincue que ce soit la seule méthode. Le RFC 7342 liste un certain nombre d'autres techniques (pas forcément directement comparables à SARP). À noter que les approches de type "*overlay*" (RFC 7364) résolvent une partie du problème mais pas la question de la taille de la table des adresses MAC. Mais il y a les RFC 4664, RFC 925, RFC 4389 (ces deux derniers sont, à mon avis, proches de SARP), RFC 4541 et RFC 6575...