RFC 7364 : Problem Statement: Overlays for Network Virtualization

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 11 octobre 2014

Date de publication du RFC : Octobre 2014

https://www.bortzmeyer.org/7364.html

En informatique, depuis les débuts des temps, on a de plus en plus virtualisé. Le processeur, puis la mémoire, puis des machines entières. Désormais, il est donc logique qu'on cherche à virtualiser le réseau. Par exemple, est-il possible de fournir, dans un centre d'hébergement de données, plusieurs centres de données virtuels, chacun complètement isolé des autres, et permettant de déployer, non pas une seule machine mais tout un réseau ? Il n'y a pas encore de solution technique complète pour ce problème mais ce RFC fournit au moins une description détaillée du problème.

Les organisations qui géreront ces centres de données virtuels seront distinctes, peut-être concurrentes. Il est donc essentiel que les différents réseaux soient parfaitement isolés les uns des autres, avec des espaces d'adressage distincts et des communications uniquement faites de manière explicite (comme s'ils étaient deux réseaux physiques connectés par un routeur). Pas question, par exemple, qu'une de ces organisations puisse "*sniffer*" le trafic d'une autre.

Le but est de gagner en souplesse d'administration. Installer un « "*data center*" » coûte cher, prend du temps, et l'infrastructure physique ne peut pas être facilement modifiée ensuite. En virtualisant tout, on pourrait imaginer que le gérant du centre de données physique loue des centres de données virtuels à différents acteurs, qui à leur tour loueront des services hébergés dans ces centres virtuels. On aura ainsi du virtuel sur du virtuel sur...

Si la virtualisation est bien faite, le locataire n'aura même pas besoin de savoir quelle est la technique sous-jacente utilisée.

Bon, c'est très joli, comme projet, mais est-ce réaliste ? Notre RFC se concentre sur la première étape, cerner le problème et poser les questions auxquelles il va falloir répondre. C'est le rôle du groupe de

travail IETF nommé NVO3 <http://tools.ietf.org/wg/nvo3>, et ce RFC, avec le RFC 7365 [1], est sa première publication. Le chiffre 3 dans le nom fait référence au fait que le groupe se concentre sur les solutions de virtualisation mises en œuvre dans la couche 3 (que ces solutions fournissent un service de couche 3 ou de couche 2). Un autre document, le RFC 7365, spécifie notamment le vocabulaire utilisé. Notre RFC y ajoute le concept de réseau supérieur ("*overlay network*"), qui désigne le réseau virtuel bâti au-dessus du réseau physique. Un exemple réalisé dans la couche 2 est 802.1Q qui fait de l'encapsulation de trames de niveau 2 dans d'autres trames.

Donc, quels sont les problèmes à résoudre, se demande la section 3 de notre RFC ? On s'attend à pouvoir déplacer les machines virtuelles d'un serveur à l'autre. En cas de virtualisation du réseau lui-même, cette mobilité doit être possible sans tenir compte de la position physique des serveurs. La machine virtuelle doit pouvoir garder son adresse IP (sinon, cela casserait les sessions TCP en cours, sans parler de la nécessité de mettre à jour le DNS, et autres informations), mais aussi son adresse MAC. En effet, des licences logicielles peuvent être liées à cette adresse MAC. En outre, d'autres perturbations pourraient surgir (système d'exploitation qui ne gère pas correctement un changement d'adresse MAC qu'il n'a pas décidé lui-même).

Dans le centre de données classique, l'adresse IP d'un serveur dépend de sa position. Par exemple, si on a un sous-réseau (et donc un routeur) pour une rangée d'armoires, préserver l'adresse IP impose de ne pas changer de rangée. Ce n'est pas une contrainte trop grave pour un serveur physique, qu'on déplace rarement. Mais ce serait très pénible pour des machines virtuelles, dont l'intérêt est justement la souplesse de placement.

Pour permettre un placement des machines à n'importe quel endroit, il est tentant d'avoir un réseau complètement plat, sans routeurs intermédiaires, ou bien en annonçant des routes spécifiques à une machine (/32 en IPv4) dans l'IGP. Mais cette absence de partitionnement, jointe à l'augmentation du nombre de machines que permet la virtualisation, met une sacrée contrainte sur les tables qu'utilisent les machines pour stocker les coordonnées de leurs voisins (ARP ou NDP en couche 2, l'IGP en couche 3). En couche 2, au lieu de stocker les adresses MAC de ses quelques voisines physiques, une machine devra stocker les milliers d'adresses MAC des machines virtuelles « voisines ». Sans compter le trafic ARP que cela représentera (cd. RFC 6820 ; un million de machines généreront plus de 468 240 ARP par seconde soit l'équivalent de 239 Mb/s).

L'exigence de souplesse et d'agilité, afin de permettre l'utilisation optimale du matériel, nécessite de séparer la configuration du réseau physique et celle du réseau logique. Le RFC donne l'exemple d'un centre de données virtuel qui serait mis en œuvre sur un petit groupe de machines physiques proches. Puis le client veut agrandir son centre et il n'y a pas de place aux alentours, on doit donc lui allouer des machines physiques situées relativement loin dans le centre physique. Cela doit pouvoir marcher, indépendamment du fait que le centre de données du client est désormais réparti en deux endroits.

Autre exigence, que les adresses utilisées sur un centre de données virtuel ne dépendent pas de celles utilisées sur un autre centre virtuel présent sur la même infrastructure physique. Dit autrement, deux clients doivent pouvoir numéroter leurs centres comme ils veulent, sans se coordonner entre eux, et sans tenir compte du plan d'adressage du réseau physique sous-jacent.

Ces exigences sont fort ambitieuses, surtout si on les combine avec la demande que la transmission des paquets suive un chemin optimal. Car la virtualisation, celle du réseau comme celle d'autres ressources, a toujours un coût en termes de performance. Par exemple, lorsqu'une machine virtuelle

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7365.txt

change de place, le meilleur routeur par défaut peut être différent, alors que cette machine n'est pas consciente de son déplacement et de la nécessité de changer. Autant que possible, l'infrastructure sous-jacente devrait faire cela pour elle. Et encore, un routeur IP classique ne garde pas d'état. Mais s'il y avait des "*middleboxes*" à état sur le trajet, comme des pare-feux ?

Pour résumer les services attendus du système de virtualisation : fournir un espace d'adressage isolé, et séparer les paquets de ceux des autres clients (on ne veut pas qu'un tcpdump sur une machine virtuelle montre les paquets d'un autre client). La section 4 décrit la classe de solutions envisagée, celle des réseaux virtuels bâtis au-dessus (d'où le terme d'"*overlay*") des réseaux physiques. Chaque réseau virtuel (pour le centre de données virtuel d'un client) serait un réseau supérieur. Lorsqu'un paquet doit être transmis, il sera encapsulé et voyagera dans un tunnel à l'extrémité duquel il sera décapsulé (technique dit "*map and encap*"). Le point d'entrée du tunnel devra faire l'opération de recherche de correspondance ("*map*"), indiquant, pour une adresse de destination donnée, quel sera le point de sortie du tunnel. Ce point d'entrée encapsulera ensuite ("*encap*") en mettant en adresse de destination de l'en-tête externe, l'adresse du point de sortie du tunnel. Comme le groupe de travail NVO3 ne considère que les réseaux virtuels mis en œuvre sur la couche 3, les adresses externes (celles de l'en-tête externe du paquet) seront forcément des adresses IP. Les adresses internes (celles du paquet avant encapsulation, ou après décapsulation) peuvent être de couche 2 ou de couche 3.

Il faudra aussi, dans bien des cas, connecter le réseau virtuel à l'extérieur (notamment à l'Internet). Cela nécessitera un routeur connecté à l'extérieur qui, après la décapsulation, transmettra le paquet par les procédés habituels (transmission IP normale, dans le cas d'une connexion à l'Internet). Même chose pour connecter deux réseaux virtuels entre eux. Si ces réseaux fournissent un service de couche 3, c'est de l'interconnexion IP classique, via un routeur.

D'un point de vue plus quantitatif, le RFC note (section 4.4) que ce système de réseaux virtuels devra fonctionner dans un environnement de grande taille (des milliers de commutateurs et des dizaines, voire des centaines de milliers de machines virtuelles) et devra donc bien passer à l'échelle. De plus, le réseau virtuel pourra être très dispersé au sein du réseau physique, avec ses fonctions mises en œuvre sur des tas de machines physiques relativement éloignées, qui travailleront aussi pour beaucoup d'autres réseaux virtuels. Le tout sera sans doute très mouvant (un des buts de la virtualisation est la souplesse, la capacité à changer souvent).

Dans le futur, on construira peut-être des centres de données physiques prévus depuis le début uniquement pour héberger des réseaux virtuels et conçus donc à cet effet. Mais, pour l'instant, ce n'est pas le cas et le RFC demande donc aussi que les solutions de virtualisation soient déployables de manière incrémentale, sans exiger une refonte complète de tout le réseau. Ainsi, les routeurs IP qui sont sur le trajet d'un tunnel ne devraient pas avoir à en être conscients. Pour eux, les paquets du tunnel sont des paquets IP comme les autres. De même, le "*multicast*" ne devrait pas être obligatoire (car il est peu déployé).

Le cahier des charges est donc gratiné. Heureusement, il n'est pas prévu que le réseau physique sous-jacent s'étende à tout l'Internet. On parle ici de virtualiser **un** centre de données, placé sous une administration unique.

Néanmoins, la réalisation de cette vision grandiose va nécessiter du travail. Le RFC envisage deux secteurs de travail : celui du contrôle et celui des données. Pour le contrôle, il faudra trouver des mécanismes efficaces pour créer, gérer et distribuer les tables de correspondance (celles qui permettent au point d'entrée de tunnel de trouver le point de sortie). Cela fera peut-être l'objet d'un tout nouveau protocole. Ce mécanisme devra permettre l'arrivée et le départ d'une machine virtuelle dans le réseau supérieur en mettant à jour ces tables rapidement et efficacement.

Pour les données, il existe déjà pas mal de formats d'encapsulation et, ici, il serait sans doute mieux de les réutiliser. Le RFC décourage donc ceux qui voudraient inventer encore un nouveau format.

La section 5 est justement consacrée aux systèmes existants et qui pourraient fournir des idées, voire des réalisations déjà opérationnelles. D'abord, on peut se servir de BGP et MPLS pour faire des réseaux virtuels IP, le RFC 4364 décrit comment. Cette solution est déjà déployée en production. Sa principale limite, selon notre RFC, est le manque de compétences BGP disponibles, alors que cette solution nécessite du BGP partout. Les mêmes protocoles permettent de faire des réseaux virtuels Ethernet (RFC en cours d'écriture).

Bon, alors, autre solution, utiliser les VLAN de 802.1. Ils fournissent une virtualisation en couche 2 en présentant plusieurs réseaux Ethernet virtuels sur un seul réseau physique. Mais ils ne reposent pas sur IP donc ne sont pas utilisables si on a, par exemple, un équipement comme un routeur sur le trajet. À l'origine, on était limité à 4 096 réseaux différents sur un seul réseau physique, mais cette limite est passée à 16 millions depuis IEEE 802.1aq.

Cette dernière technique, également nommée SPB, utilise IS-IS pour construire ses réseaux virtuels.

Et TRILL ? Il utilise lui aussi IS-IS pour fournir des réseaux de couche 2 différents. En France, cette technique est par exemple déployée chez Gandi <http://www.gandibar.net/post/2013/02/05/Is-16-Million-VLANs-enough-for-you>.

Plus exotique est le cas de LISP (RFC 9300). Son but principal n'est pas de virtualiser le "*data center*" mais il pourrait contribuer, puisqu'il fournit des réseaux virtuels IP au-dessus d'IP (donc, service de couche 3, mis en œuvre avec la couche 3).

Enfin, bien que non cité par le RFC, il faudrait peut-être mentionner VXLAN, bien décrit en français par Vincent Bernat <http://vincent.bernat.im/fr/blog/2012-multicast-vxlan.html> et en anglais par le Network Plumber <http://linux-network-plumber.blogspot.fr/2012/09/just-published-linux-kernel.html>.

À noter qu'un groupe de travail de l'IETF, ARMD <http://tools.ietf.org/wg/armd>, avait travaillé sur un problème proche, celui posé par la résolution d'adresses IP en adresses MAC (avec ARP ou NDP) dans les très grands centres de données où tout le monde était sur le même réseau de couche 2. Le résultat de leurs travaux avait été documenté dans le RFC 6820.

En résumé, notre RFC conclut (section 6) que le problème est complexe et va nécessiter encore du travail...

Merci à Nicolas Chipaux et Ahmed Amamou pour leur relecture.