RFC 7365 : Framework for Data Center (DC) 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/7365.html

Ce nouveau RFC décrit le cadre général de la **virtualisation de réseaux** en utilisant IP comme substrat. Il est issu du projet NVO3 <https://tools.ietf.org/wg/nvo3> de l'IETF, qui a produit une description du problème (dans le RFC 7364 [1]) et ce document. À ce stade, ce n'est pas encore un protocole concret, même s'il existe déjà des solutions techniques partielles pour ce problème.

La cible est constituée des grands centres de données, hébergeant des centaines de milliers de machines virtuelles, et qu'on souhaite pouvoir gérer de manière souple (par exemple en déplaçant facilement une machine virtuelle d'une machine physique à une autre, sans l'éteindre et sans la changer d'adresse IP). Le RFC 7364 décrit plus en détail le problème à résoudre (et il est donc recommandé de le lire d'abord), ce RFC 7365 étant le cadre de(s) la(les) solution(s). Comme on envisage des grands ensembles de machines, et qu'on cherche à être très dynamique (reconfigurations fréquentes), il faudra beaucoup d'automatisation (pas question de modifier des tables à la main parce qu'une machine s'est déplacée).

Il définit donc un vocabulaire et un modèle pour ces réseaux virtuels. Donc, on parle de **réseaux NVO3** ("*NVO3 networks*"). Il s'agit de créer des réseaux virtuels (VN pour "*virtual networks*", ou bien "*overlays*") fonctionnant au-dessus d'un protocole de couche 3, IP. Ces réseaux virtuels fournissent aux clients (qui peuvent appartenir à des organisations distinctes, voire concurrentes) soit un service de couche 3, soit un service de couche 2 . Les réseaux virtuels sont créés au-dessus d'un substrat, l'"*underlay*", qui est le « vrai » réseau sous-jacent. Le substrat est toujours de couche 3, comme le nom du projet NVO3 l'indique. Chaque VN aura un **contexte**, un identificateur qui figurera dans les paquets encapsulés et qui permettra à l'arrivée de distribuer le paquet au bon VN. Le réseau NVO3 utilisera des NVE ("*Network Virtualization Edge*") qui seront les entités qui mettent en œuvre ce qui est nécessaire pour la

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7364.txt

virtualisation. Un NVE a au moins une patte vers le client (qui envoie des paquets sans savoir qu'il y a virtualisation) et une autre vers le réseau IP qui transporte le trafic entre les NVE. Un NVE n'est pas forcément un équipement physique, cela peut être, par exemple, le commutateur virtuel d'un hyperviseur.

Les sections 2 et 3 présentent le modèle de référence utilisant ce vocabulaire, et ses composants. Les systèmes des clients ("*tenant systems*") sont connectés aux NVE, via un VAP ("*virtual access point*", qui peut être une prise physique ou bien virtuelle, par exemple un VLAN ID). Un même NVE peut présenter aux clients plusieurs VNI ("*virtual network instance*", une instance particulière de réseau virtuel). NVE et système du client peuvent être physiquement dans la même boîte, ou bien il peut s'agir de systèmes distants (le schéma 2 du RFC n'est pas une représentation physique, mais logique). Dans le premier cas, la communication de l'information peut se faire par une API locale, dans le second, par un protocole réseau. Les NVE échangent l'information entre eux (comme des routeurs avec un IGP) ou bien ils sont connectés à une NVA ("*Network Virtualization Authority*") qui leur fournit les informations sur l'état du réseau (qui est membre, comment joindre les membres, etc). Il y aura donc un protocole entre NVE, ou bien entre NVE et NVA. Le NVA n'est évidemment pas forcément un serveur unique, cela peut être une grappe, pour la redondance. Les machines qui forment l'"*underlay*" font du routage IP normal et ne connaissent pas, a priori, les systèmes des clients. La gestion de ce réseau "*underlay*" se fait avec les mêmes outils OAM que n'importe quel réseau IP.

Le réseau virtuel (l'"*overlay*") devra donc utiliser une forme ou l'autre d'encapsulation pour faire passer ses paquets sur l'"*underlay*". Cela pourra être GRE, IPsec, L2TP, etc. Ce n'est pas tout de tunneler, il faut aussi un mécanisme de contrôle, rassemblant et distribuant l'information dans le réseau. Un tel mécanisme peut être centralisé (comme dans le cas de SDN) ou réparti (comme l'est traditionnellement le routage dans l'Internet).

Quant un NVE fournit un service de couche 2, les systèmes des clients ont l'impression de se connecter à un Ethernet normal (comme avec le RFC 4761 ou le RFC 4762). S'il fournit un service de couche 3, les systèmes des clients voient un réseau IP, comme avec le RFC 4364.

Autre service important, la possibilité de déplacer des VM à l'intérieur du centre de données. Permettre des déplacements à chaud, sans arrêter la VM, est évidemment très souhaitable, mais cela aura des conséquences sur des points comme les caches ARP (qu'il faut mettre à jour si une VM se déplace, alors qu'elle pense rester dans le même réseau de niveau 2).

La section 4 du RFC décrit quelques aspects essentiels dans les discussions sur les réseaux virtuels. D'abord, les avantages et inconvénients des réseaux virtuels. Sur le papier, ils offrent des tas d'avantages : les systèmes des clients n'ont à s'occuper de rien, ils bénéficient d'un réseau certes virtuel mais qui a les mêmes propriétés qu'un vrai, tout en étant plus souple, avec des changements plus rapides. Même les adresses (MAC et IP) sont séparées de celles de l'"*underlay*" donc les clients sont isolés les uns des autres. Malheureusement, il y a quelques problèmes en pratique : pas de contrôle sur le réseau sous-jacent et même pas d'information sur lui, par exemple sur son utilisation, ou sur des caractéristiques comme le taux de perte de paquets (RFC 7680). Il y a donc un risque de mauvaise utilisation des ressources réseaux. Et si plusieurs "*overlays*" partagent le même "*underlay*", l'absence de coordination entre les réseaux virtuels, chacun ayant l'impression qu'il est tout seul (« "*each overlay is selfish by nature*" »), peut encore aggraver les choses. Mais, bon, des réseaux virtuels existent depuis de nombreuses années, et marchent.

Autre exemple, parmi ceux cités par le RFC, des difficultés à réaliser proprement ce projet, le cas de la diffusion. On veut pouvoir fournir des services qui existent sur les réseaux physiques, comme la capacité à diffuser à tous les membres du réseau, avec un seul paquet. Comment fournir le même service avec des réseaux virtuels ? Il va falloir automatiquement répliquer le paquet pour le transmettre ensuite à tous les

réseaux physiques qui contribuent au réseau virtuel. Cela peut se faire de manière violente (dupliquer le paquet en autant d'exemplaires qu'il y a de machines et le transmettre à chacune en "*unicast*") ou bien de manière plus subtile mais plus complexe avec le "*multicast*". La première solution minimise l'état à conserver dans les routeurs, la seconde minimise la capacité réseau consommée.

La mise en œuvre concrète de la virtualisation va nécessiter des tunnels, connectant les différents réseaux physiques. Qui dit tunnel dit problèmes de MTU, en raison des octets consommés par les entêtes du format de tunnel utilisé. Ces quelques octets en plus diminuent la MTU du réseau virtuel. En théorie, la fragmentation résout le problème mais elle a un prix élevé en performances (idem si fragmentation et réassemblage étaient faits dans le système de virtualisation et non pas, comme la fragmentation IP habituelle, par les machines terminales et - en IPv4 - les routeurs). La découverte de la MTU du chemin (RFC 1981 et RFC 4821) permet de se passer de fragmentation mais, en pratique, elle marche souvent mal. Il semble que la meilleure solution soit de faire en sorte que le réseau virtuel présente une MTU « habituelle » (1 500 octets, la MTU d'Ethernet) et, pour réaliser cela, que la MTU de l'"*underlay*" soit assez grande pour que les octets de l'encapsulation puissent être ajoutés sans problèmes (utilisation de jumbogrammes).

Enfin, les réseaux virtuels posent des problèmes de sécurité spécifiques (section 5). Un système client ne doit pas pouvoir attaquer le réseau sous-jacent (il ne doit même pas le voir), et un système client ne doit pas pouvoir attaquer un autre système client (par exemple, s'ils sont dans des réseaux virtuels séparés, il ne doit pas pouvoir lui envoyer directement de paquet, même si les deux systèmes clients se trouvent être sur la même machine physique).

Du point de vue du client, le critère est la capacité du réseau NVO3 à distribuer son trafic au bon endroit, et à séparer les trafics. Un tcpdump chez un client ne doit pas montrer les paquets d'un réseau virtuel d'un autre client (ce qui ne dispense évidemment pas d'utiliser le chiffrement, pour les attaques effectuées par le réseau sous-jacent). Et les machines doivent pouvoir faire confiance au VNI (identificateur du réseau virtuel) indiqué dans un paquet. Comme avec le RFC 2827, une machine client ne doit pas pouvoir injecter des paquets mentant sur le réseau virtuel auquel elle appartient.

Des futurs protocoles conformes à ce schéma sont en cours d'examen dans le groupe de travail `<http://tools.ietf.org/wg/nvo3/>`, mais qui seront en concurrence avec des protocoles déjà existants qui ont les mêmes objectifs, comme celui du RFC 4364.

Merci à Thomas Morin et Marc Lasserre pour leur relecture.