Nouveau type de VLAN, le VXLAN

Introduction

Il est aujourd'hui monnaie courante pour les hébergeurs et les opérateurs Telco de fournir des réseaux virtuels sur lesquels leurs clients peuvent déployer leurs VLANs 802.1Q.
En raison du nombre limité de VLANs possible (4096) en 802.1Q , les constructeurs ont mis au point de nouvelles technologies de VLANs déployable dynamiquement au dessus d'un réseau IP routé.

VxLAN est l'une d'entre elles.

Nous sommes là au coeur des technologies réseau permettant de faire fonctionner des clouds répartis sur plusieurs DataCenters.

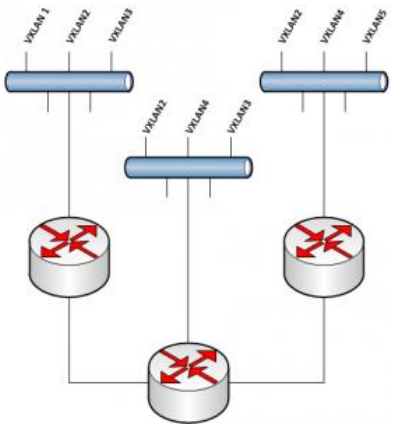
Présentation de VxLAN

VxLAN est un format d'encapsulation porté par Cisco et VMware ayant des fonctionnalités semblables aux VLANs mais avec quelques améliorations que nous allons détailler dans cet article.

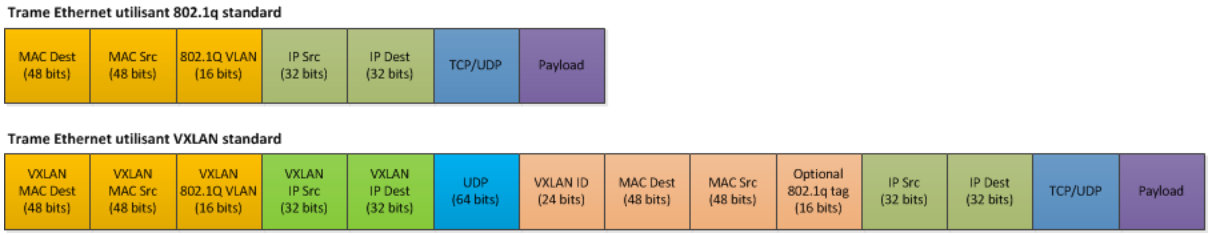
VxLAN est l'acronyme de Virtual eXtensible LAN est une standardisation à été proposée à l'IETF en 2011.

Grâce à ce format, il est possible de faire transiter des trames de niveau 2, dans UDP.

Cette propriété permet, par conséquent, d'utiliser une segmentation de type VLAN au delà d'un domaine Ethernet.



De plus, grâce à son champ d'identification sur 24 bits, il est possible de créer sur un même domaine VxLAN plus de 16 millions de VLANs différents (224 plus précisément). Cependant, puisque VxLAN est un format d'encapsulation, il provoque une surcharge d'entête dans les trames ethernet. Le schéma ci-dessous compare une trame Ethernet avec VLAN "standard" avec une trame Ethernet avec VxLAN.



Comme on peut le constater sur ce schéma, VxLAN provoque un "overhead" (i.e surcharge d'encapsulation) assez important (A peu de chose près équivalente à la taille des entêtes de niveau 2 + niveau 3 + niveau 4) soit environ 50 octets.

Un peu de vocabulaire

ARP : *Address Resolution Protocol*, mécanisme de base permettant d'obtenir l'adresse physique d'une machine en fonction de son IP. Protocol de base sur lequel repose IP.

MTU : *Maximum Transmission Unit*, correspond à la taille maximum transmissible sur une interface (Ethernet dans notre cas) sans fragmentation.

Multicast : Méthode de diffusion permettant à une machine de communiquer avec un groupe de machines avec un système d'abonnement

TTL : *Time To Live*, valeur incluse dans un paquet IP étant décrementé lors du passage dans un routeur permettant d'éliminer les paquets entrés dans des boucles de routage.

VTEP : *VXLAN Tunnel End Point*, correspond à la porte d'entrée (ou de sortie) d'un domaine VXLAN : Ces éléments permettent l'encapsulation et la désencapsulation des paquets VXLAN afin d'acheminer ceux-ci vers leur destination finale. Cette fonctionnalité peut-être assuré par le noyau Linux depuis le noyau 3.7.

Principe de fonctionnement

VxLAN repose sur l'encapsulation de trames Ethernet dans UDP.

Le schéma simplifié d'une trame réseau utilisant VXLAN est donc le suivant :

Niveau 5	MAC Dest	MAC Src	IP Src	IP Dest	TCP/UDP	DATA
Niveau 4	UDP					
Niveau 3	IP VTEP Source	IP VTEP Destination				
Niveau 2	MAC VTEP Dest (ou MAC GW)	MAC VTEP Source				
Niveau 1	Cuivre / Fibre					

On voit ci-dessus que le paquet originellement emis par la machine dans le domaine VxLAN ne se retrouve qu'au niveau 5. Il est donc possible d'effectuer un routage entre les VTEP avant de décapsuler le paquet et ainsi permettre une segmentation de type V(X)LAN en s'affranchissant des domaines Ethernet.

Etudions maintenant comment fonctionnent les fonctions de niveau 2 (ARP, broadcast, …) au sein d'un domaine VxLAN.

Alors que sur un domaine Ethernet standard, ces protocoles fonctionnent grâce à l'adresse de broadcast, un domaine VXLAN utilise quand à lui une adresse IP Multicast.

De cette façon, lors du déploiement d'un nouveau VTEP sur un domaine VxLAN, il faudra simplement l'abonner au groupe multicast correspondant à son VxLAN.

VXLAN dispose ensuite d'un mecanisme faisant transiter les données de broadcast sur ce groupe multicast permettant à tous les VTEP de communiquer ensemble afin de répondre (par exemple) aux requêtes ARP.

Ceci ce traduit lors de la création d'une interface vxlan, par une commande du type :

```
$ ip link add vxlan10 type vxlan id 10 group 239.0.0.10 ttl 10 dev eth0

$ ip link set vxlan10 up
```

Grâce à ces deux commandes nous créons une interface VxLAN avec le tag 10 dont l'interface physique sera eth0, dont le domaine de broadcast est géré grâce à l'adresse multicast 239.0.0.10 et le TTL des packets encapsulés à 10.

De ce fait, toutes les machines abonnées au groupe multicast 239.0.0.10 recevront le trafic de broadcast de ce VXLAN.

Limites et remarques

Le fonctionnement de VxLAN reposant sur IP, celui-ci donne aux architecture l'utilisant l'avantage d'être scalable. Seules les limites des VTEP et de la bonne configuration de votre réseau peuvent limiter les performances du protocole.

En effet, comme nous l'avons vu dans l'introduction, ce protocole d'encapsulation provoque une surcharge d'entête. Celle ci pourrait avoir des conséquences importantes sur les performances de votre réseau si celui-ci est mal configuré.

Plus précisement, avec une surcharge d'environ 50 octets en IPv4 et 100 octets en IPv6, votre VTEP est susceptible d'avoir à générer des trames Ethernet de 1550 à 1600 octets. Si votre réseau est configuré avec une MTU à 1500 (comme dans la majorité des cas), vous risquez d'avoir un taux de fragmentation assez important qui aura un impact sur les performances du réseau.

Afin de résoudre ce problème, deux solutions existent :

1) Limiter vos machines dans un domaine VXLAN à envoyer des trames plus petite (MTU à 1450 ou 1400o)

2) Augmenter le MTU global de votre réseau à 1550 ou 1600, si vous pouvez le paramétrer de bout en bout.

Source : http://www.randco.fr/actualites/2013/vxlan/