



SERVICE DE VPN DE NIVEAU 2 SUR RAP : VPLS

Description : Ce document présente l'architecture VPLS/MPLS pour la fourniture du service VPN de niveau 2 sur RAP.

Version actuelle : 1.3

Date : 10/04/09

Auteur : LD

Version	Dates	Remarques
1.0	01/04/09	Création du document
1.1	07/04/09	Ajout des parties 4,6,7,8
1.2	08/04/09	Réorganisation sous-parties
1.3	10/04/09	Relecture LGY

TABLE DES MATIÈRES :

Table des matières :	2
I. Contexte :	3
II. Architecture mpls de rap : plan de données pour la prise en charge de vpls.....	3
a. Construction d'un nuage MPLS :	3
b. Optimisation du nuage MPLS : Redondance et Haute-disponibilité	4
III. Mise en œuvre de VPLS :	6
a. Plan de contrôle de VPLS : Signalisation et auto-decouverte BGP	6
b. Plan de données pour la commutation dans un domaine VPLS :	7
IV. Intégration d'un site dans un domaine vpls :	8
V. Fonctionnement de vpls :	9
VI. VPLS multihoming :	10
VII. Fast-convergence :	11
VIII. Classes de services :	11
IX. Bibliographie :	11
X. Glossaire :	11

I. CONTEXTE :

Avec le déploiement de sa nouvelle infrastructure, le Réseau Académique Parisien a fait le choix de mettre en œuvre MPLS dans le but de fournir un service de VPN de niveau 2 basé sur VPLS.

Auparavant fourni par la propagation de VLAN 802.1Q sur son backbone, le service VPN de niveau 2 sur RAP s'est avéré de plus en plus contraignant à opérer, par le nombre de tags limité, par la nécessité de garantir l'unicité d'un tag sur un chemin traversant plusieurs réseaux autonomes et par les difficultés d'exploitation pouvant être rencontrées dans l'utilisation du protocole de Spanning-Tree en cas de bouclage.

La fourniture de ce service a donc évolué grâce au protocole VPLS, basé sur la nouvelle infrastructure MPLS et chaque VLAN 802.1Q a été migré vers une instance VPLS. Cela permet d'offrir un service de VPN de niveau 2 de haute fiabilité (redondance et fast-convergence) sur le backbone ainsi que sur l'accès des sites en raccordement fiabilisé sur RAP avec son mécanisme de multihoming.

II. ARCHITECTURE MPLS DE RAP : PLAN DE DONNEES POUR LA PRISE EN CHARGE DE VPLS.

La mise en œuvre de MPLS permet de réaliser un routage et une commutation efficace ainsi que de déployer des services à fortes valeurs ajoutées : VPN, VPLS, Ingénierie de trafic.

a. CONSTRUCTION D'UN NUAGE MPLS :

Pour construire un nuage MPLS, il faut créer des tunnels MPLS (*Pseudo-wire* : *PW*) entre chacun des PE du réseau, chacun des tunnels étant composé de 2 LSP unidirectionnels comportant un label de commutation, le label MPLS. Un LSP est établi entre 2 PE, le PE d'entrée : *ingress router* et le PE de sortie : *egress router*. Si l'on crée un LSP entre 2 PE non adjacents (ce qui est le cas sur RAP où il existe d'une part un niveau 2 entre chaque PE adjacent et d'autre part un niveau 2 entre chaque PE et son PE d'accès à l'extérieur (PE-Jussieu et PE-Odeon)), alors il y aura un PE de transit sur le chemin de ce LSP.

i. SIGNALISATION RSVP-TE :

Pour mettre en œuvre ces tunnels, il existe 3 méthodes : statique, LDP ou RSVP. Le choix s'est porté sur cette dernière pour la mise en place d'une signalisation entre tous les PE, car elle permet de créer de manière automatique un maillage complet de LSPs. Mais RSVP ne se limite pas à cette fonction, ce protocole présente aussi des avantages en terme de haute disponibilité et potentiellement de réservation de ressource (même si ce point n'est pas à l'ordre du jour sur

RAP), il faut d'ailleurs parler de RSVP-TE dans notre cas, car il s'agit d'une extension du protocole RSVP.

ii. MODE D'ETABLISSEMENT DES LSP :

Il existe plusieurs modes d'établissement automatique des LSP avec RSVP-TE et c'est celui qui se base sur les informations de topologie disponibles grâce à l'algorithme CSPF qui a été choisi. Cet algorithme (extension du protocole OSPF : OSPF-TE *Traffic Engineering*) tient à jour une table de routage TED (*Traffic Engineering Database*) qui, en plus de la topologie du réseau (que l'on retrouve de manière basique avec la RIB de l'IGP) tient compte d'informations telles que la charge des liens, le taux de ressources disponibles sur les routeurs d'un chemin, etc... Dans le cas de RAP, seule la topologie du réseau est utilisée pour l'établissement des LSP, mais l'utilisation de CSPF est nécessaire dans le cas d'un mode d'établissement dynamique des LSP avec RSVP-TE. Pour simplifier, on peut dire que chaque LSP est donc établi entre 2 PE par le chemin dont le coût OSPF est le moins élevé :

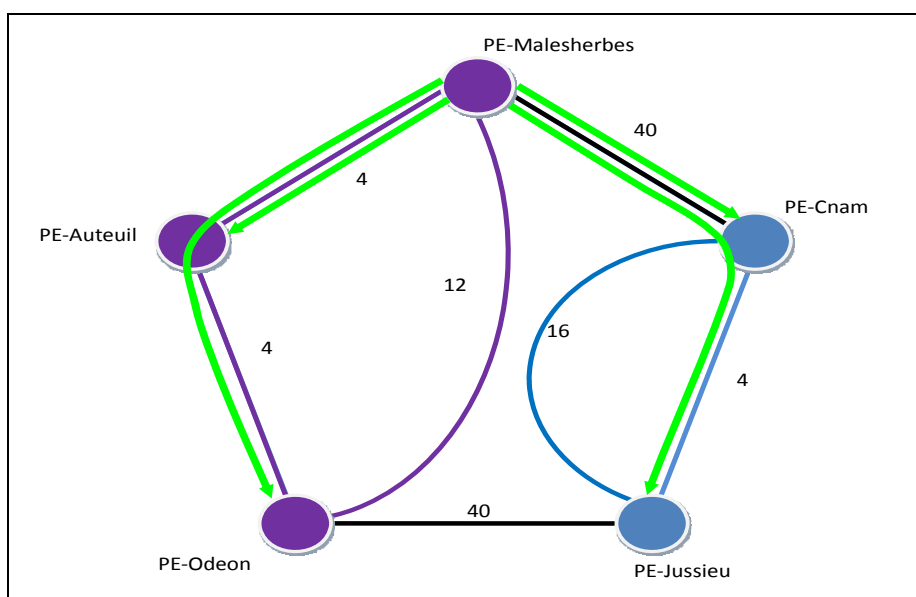


FIGURE 1 : ETABLISSEMENT DE LSP AVEC RSVP-TE ET CSPF (COUTS OSPF), EX. PE-MALESHERBES

b. OPTIMISATION DU NUAGE MPLS : REDONDANCE ET HAUTE-DISPONIBILITE

Pour redonder les LSP, on fait appel au mécanisme « Standby Secondary paths». Il s'agit de configurer un second LSP qui sera le backup du LSP primaire entre deux PE. En cas de coupure du LSP primaire, ce LSP secondaire est déjà établi et est prêt à prendre le relais, c'est une optimisation permettant une convergence en moins d'une seconde.

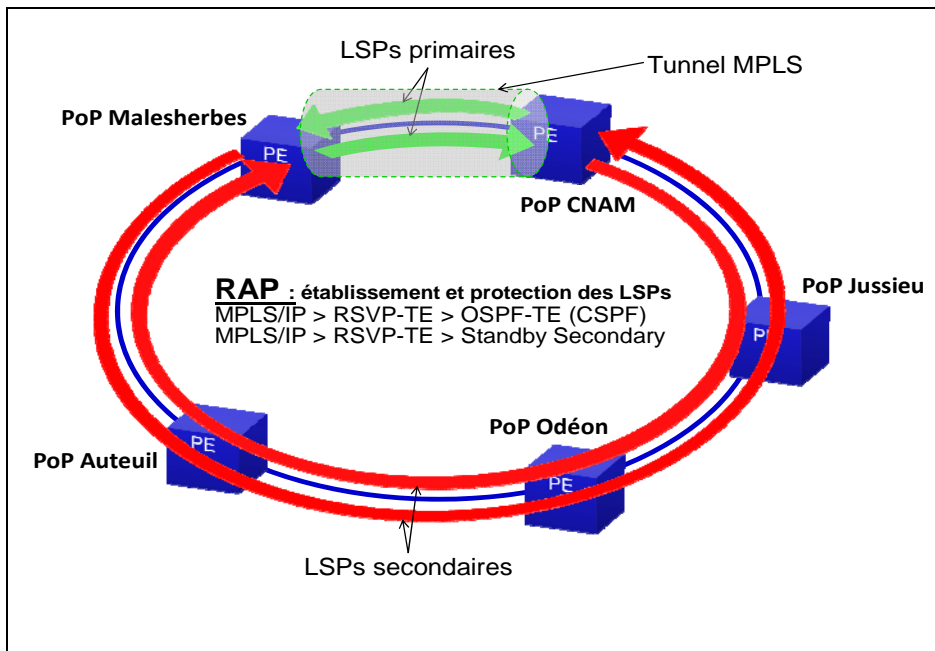


FIGURE 2 : ETABLISSEMENT DES TUNNELS MPLS

Typiquement sur RAP, un LSP secondaire fera le grand tour de la zone (Zone Odéon ou zone Jussieu) dans laquelle se situe le PE :

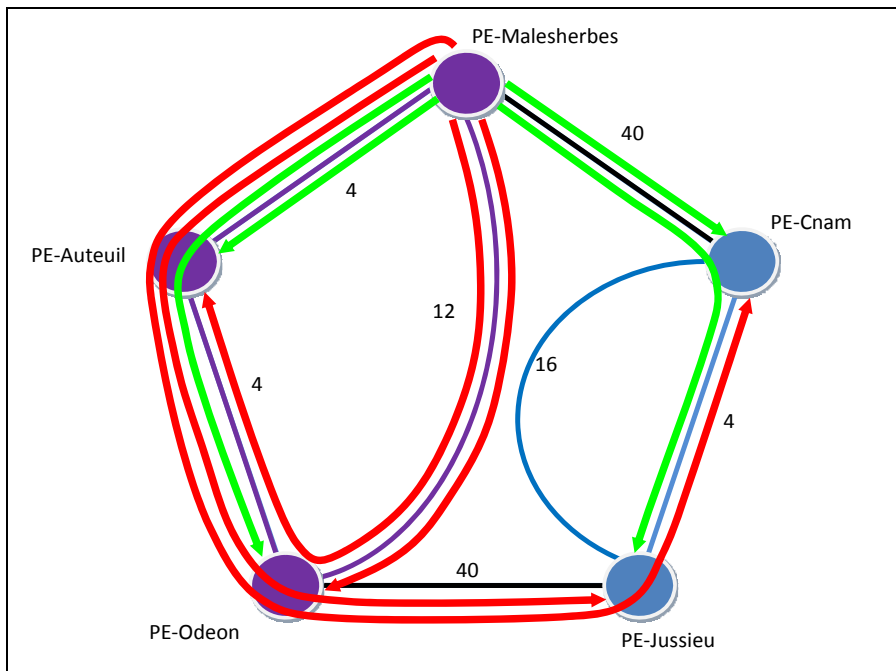


FIGURE 3 : PE-MALESHERBES, ETABLISSEMENTS DE LSP PIRMAIRES ET SECONDAIRES AVEC CHACUN DES AUTRES PE EN FONCTION DE CSPF

Ce premier niveau permet d'offrir une continuité de service, mais pour un LSP donné, c'est sur le routeur *ingress* du LSP que s'effectue le changement de LSP. Ainsi, un autre mécanisme permet d'optimiser les temps de convergence en cas de coupure d'un lien en permettant à un routeur de transit (qui se situe sur le chemin d'un LSP donné, entre le routeur *ingress router* et le routeur *egress router* du nuage MPLS) de rerouter le trafic, c'est le mécanisme de *fast reroute*. Chaque *P router* du réseau (les PE dans le cas de RAP) maintient des « détours » pour chaque LSP dont il fait partie, pour joindre chaque PE voisin en cas de coupure de lien à n'importe quel endroit du réseau. *Fast reroute* est aussi une extension du protocole RSVP-TE et permet donc, le temps de commuter sur le LSP secondaire, à ce que les paquets déjà partis dans le LSP primaire puissent être reroutés par les *P router* de transit.

III. MISE EN ŒUVRE DE VPLS :

La technologie VPLS offre un service Ethernet multipoint à multipoint qui apporte une connectivité entre plusieurs sites et simule, de manière transparente, une liaison LAN Ethernet entre chacun des sites. On parle ici de domaine ou *instance* VPLS. Il existe deux approches pour l'implémentation de VPLS et c'est la solution « Draft Kompella¹ » qui a été mise en œuvre, basée sur BGP en termes de signalisation et de découverte automatique, l'autre cas étant l'utilisation de LDP (ce dernier ne permet que la signalisation).

a. PLAN DE CONTROLE DE VPLS : AUTO-DECOUVERTE ET SIGNALISATION BGP

Pour mettre en œuvre VPLS, tous les PE ont un peering iBGP avec les 2 routeurs *route-reflector* que sont les PE-ODEN et PE-JUSSIEU. Ceci est déjà le cas pour l'annonce des routes pour le trafic IP mais ici on utilise la « BGP VPLS NLRI » : famille d'adresse « I2vpn » (AFI L2VPN) avec la sous-famille VPLS (SAFI VPLS).

¹ Le terme est resté, mais il s'agit maintenant du RFC4761 : *Virtual Private LAN Service (VPLS) using BGP for Auto-discovery and Signaling*.

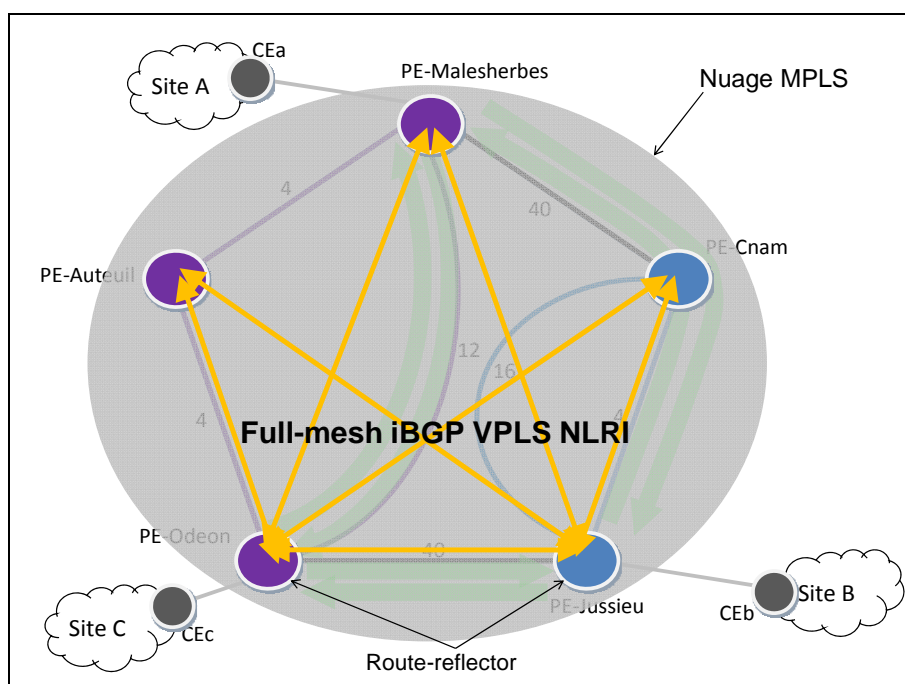


FIGURE 4 : PEERING iBGP VPLS NLRI ENTRE CHAQUE PE ET LES ROUTE-REFLECTOR

Ces annonces BGP permettent l'auto-découverte par un PE des PE appartenant² à un domaine VPLS donné : un PE annonce qu'il appartient à un domaine VPLS en intégrant dans ses annonces BGP NLRI VPLS une *Route Target*³. Cela permet à tous les PE appartenant à ce domaine VPLS d'obtenir toutes les informations nécessaires en vue d'établir des LSP (Ici, on parle de LSP dédiés à un domaine VPLS) avec ce nouveau PE de manière automatique. Cet établissement de LSP, et donc de PW dédiés entre chacun des PE d'un domaine VPLS s'effectue grâce à la signalisation, mécanisme aussi réalisé grâce à BGP et l'annonce des informations *Route Distinguisher, label VPLS, etc*⁴.... Il ne s'agit pas ici de la création de tunnels pour la commutation de l'ensemble des paquets entre les PE (déjà établis par les LSP RSVP/MPLS), mais de la création de tunnels de second niveau, spécifiques à un domaine VPLS donné, et permettant aux PE de gérer le plan de données (commutation) pour les sites connectés à ce domaine VPLS.

b. PLAN DE DONNEES POUR LA COMMUTATION DANS UN DOMAINE VPLS :

Après l'établissement de LSP grâce à BGP, le trafic dans un domaine VPLS sera commuté selon les labels de ces LSP, qui eux-mêmes « transiteront » par les LSP du backbone mis en place par

² Un PE appartient à un domaine VPLS si un site qu'il raccorde est client de ce domaine et que l'on a configuré pour cela une interface logique.

³ Route Target : communauté étendue BGP permettant d'identifier l'instance (ou domaine) VPLS pour laquelle les informations suivantes de l'annonce BGP sont à prendre en compte.

⁴ Plus d'information dans le RFC 4761.

MPLS/RSVP-TE, ayant leur propre label aussi. Les LSP VPLS bénéficient donc des mécanismes de redondance et de fast-convergence. Il y a donc un label de commutation par LSP entre chaque paire de PE faisant partie d'un domaine VPLS.

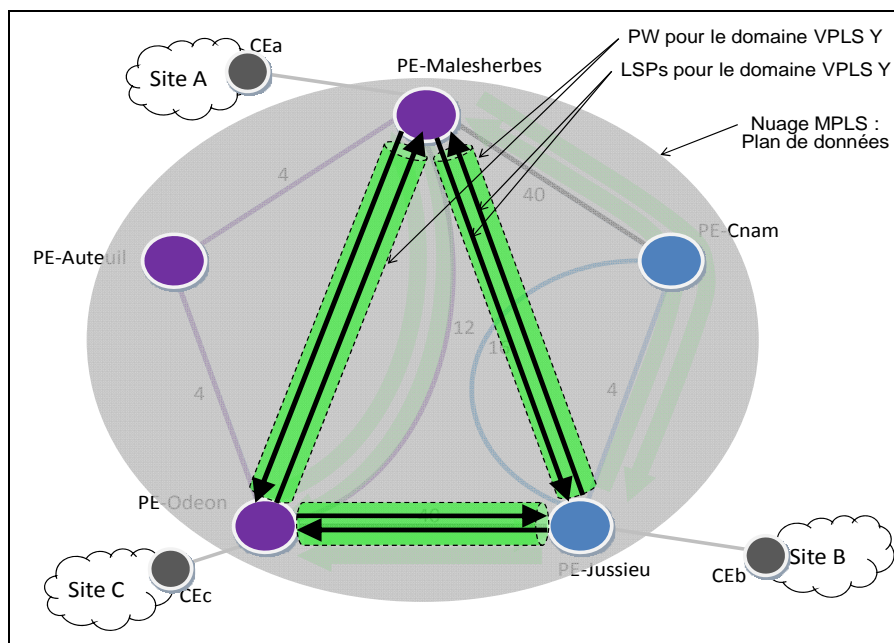


FIGURE 5 : ETABLISSEMENT DES LSPS POUR LE DOMAINE VPLS Y ENTRE LES SITES A, B ET C

IV. INTEGRATION D'UN SITE DANS UN DOMAINE VPLS :

Un CE⁵ n'a besoin d'implémenter ni MPLS ni VPLS, le PE se charge d'encapsuler (*encapsulation vlan-vpls*) les trames Ethernet 802.1Q provenant du site dans le domaine VPLS correspondant. Cette correspondance se fait grâce à la création d'une interface logique (sur le port de raccordement du site du PE) dont le tag est le numéro d'instance du domaine⁶ VPLS : un mapping est effectué entre le tag 802.1Q et le numéro d'instance VPLS : il s'agit de l'encapsulation d'une trame 802.1Q dans une trame VPLS.

Et, puisque les domaines VPLS sont totalement indépendants les uns des autres sur le backbone de RAP, le site peut choisir n'importe quel tag 802.1Q disponible sur sa liaison d'accès.

⁵ CE : Equipement d'accès de site, il n'implémente pas nécessairement MPLS.

⁶ « instance du domaine » : dans la configuration on parle d'instance VPLS, dans la technologie on parle de domaine VPLS, ici, le mot instance est à prendre hors contexte « configuration ».

V. FONCTIONNEMENT DE VPLS :

Une fois le domaine VPLS établi, un PE peut participer comme un commutateur Ethernet dans la vie de son LAN : ici, on parle de *VPLS Edge Device (VE)* ou de *Virtual Bridge (VB)* qui va de la même manière utiliser les fonctionnalités classiques que l'on retrouve dans Ethernet : MAC learning, packet replication and forwarding, etc...Il est prêt à recevoir des trames Ethernet d'un site client et peut commuter ces trames sur le LSP approprié en fonction de l'adresse MAC destination. Cela est possible car un PE tient à jour une table d'adresses MAC (FIB) par domaine VPLS.

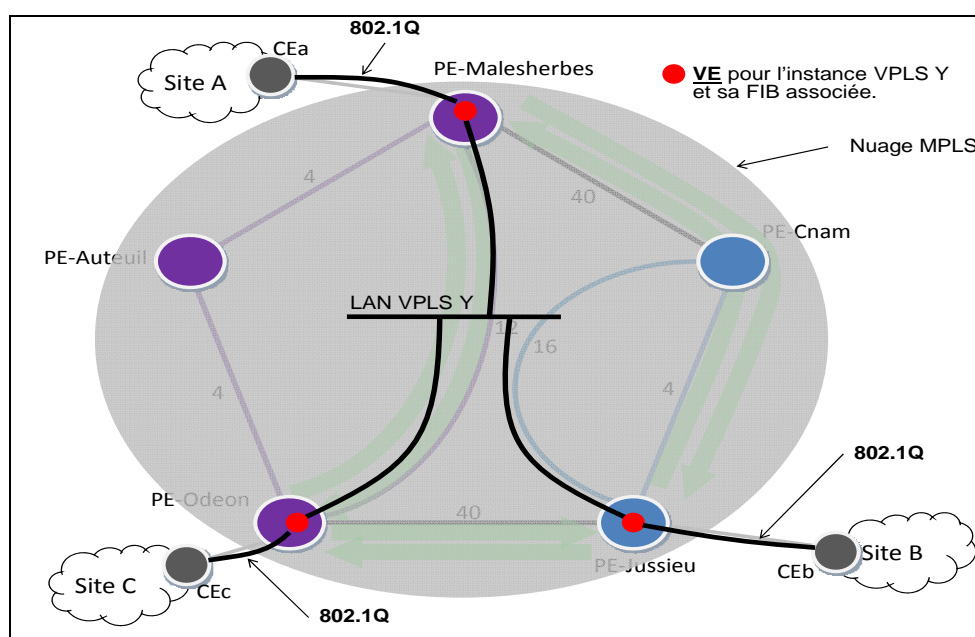


FIGURE 6 : DOMAINE VPLS, SIMULATION D'UN LAN ETHERNET

Le PE alimente donc une FIB par domaine VPLS en fonction des trames Ethernet en provenance du/des site(s), la remplit (*MAC learning*) et supprime une entrée inutilisée depuis trop longtemps (*Aging time*).

Dans une FIB VPLS, chaque adresse MAC a un LSP de correspondance, ainsi lorsqu'une trame Ethernet provenant d'un site arrive sur un PE et dont l'adresse MAC destination est connue, il lui suffit de la transmettre dans ce LSP.

Si l'adresse MAC destination n'est pas renseignée dans la FIB, la trame Ethernet est répliquée et transmise sur tous les LSP du domaine VPLS correspondant, sauf le port d'entrée de cette trame. S'en suit un *flooding*, les PE qui recevront cette trame vont à leur tour la répliquer vers

les sites qu'ils raccordent et appartenant à ce domaine VPLS, tout en évitant les boucles avec le mécanisme de *split-horizon* (un PE de sortie du domaine VPLS ne va pas répliquer la trame reçue vers les autres PE mais uniquement vers les sites qui lui sont raccordés).

VI. VPLS MULTIHOMING :

En plus des fonctionnalités présentées ci-dessus, VPLS permet, pour un site connecté à 2 PE, de redonder ses VPN de niveau 2 sur son accès à RAP. Les 2 PE en question sont alors configurés dans un même domaine VPLS, avec un identifiant identique (VE ID). Il faut alors configurer un « poids » (*site preference*) sur les 2 PE pour obtenir un PE nominal et un PE backup. La négociation se fait alors en BGP entre les 2 PE avec l'information de *preference* (de la même manière qu'une *local-preference* en BGP), on parle de *BGP Path selection*.

Sur le backbone de RAP, cela se traduit par des LSP qui sont actifs depuis le PE dont la *preference* est la plus forte.

Sur les interfaces site des PE, cela se traduit pour chaque service de niveau 2 par une interface active UP et une interface passive DOWN.

En cas d'incident sur la liaison vers le site ou sur le PE nominal, les mécanismes de redondance et de fast-convergence permettent une convergence de l'ordre de la seconde.

Cette fonctionnalité permet donc à un site de prolonger cette redondance jusqu'à l'utilisateur final, dont le réseau aura été pourvu lui-même de mécanismes le permettant : MPLS, *STP, Flex-link, VSS, etc...

Pour les sites en raccordement fiabilisé sur RAP⁷, cette fonctionnalité est activée systématiquement, tous les VLAN du site sont disponibles sur chacun des PE d'accès du site. C'est le PE nominal qui est actif pour l'ensemble des services de niveau 2 du site ; Le PE secondaire n'est actif qu'en cas de panne. Ainsi, le site doit mettre en œuvre la bonne architecture et les bons mécanismes pour que ses services de niveau 2 soient utilisables au niveau du lien de secours en cas de panne. Il doit aussi faire attention aux configurations afin d'éviter les boucles. Puisque que seules les interfaces logiques du PE nominal sont actives dans les niveaux 2 d'un site, il ne peut y avoir de concurrence ou de problème d'asymétrie entre les 2 accès du site, toutefois, le site doit rester en alerte quant aux dangers potentiels de bouclage après avoir activé ses niveaux 2 sur son routeur d'accès de backup.

⁷ Cf. le portail de RAP pour les détails techniques d'un raccordement fiabilisé aux niveaux 1 et 3.

VII. FAST-CONVERGENCE :

Comme indiqué plus haut, les temps obtenus en cas de coupure de lien ou d'incident sur un PE sont de l'ordre de la seconde. Ces performances permettent de garantir, pour la ToIP notamment, qu'une communication ne sera pas coupée en cas d'incident sur un lien ou un PE. La coupure n'est quasiment pas ressentie par l'utilisateur final.

VIII. CLASSES DE SERVICES :

Comme pour les VLAN 802.1Q, les VPN de niveau 2 en VPLS bénéficient des classes de services sur RAP avec le champ MPLS EXP.

IX. BIBLIOGRAPHIE :

- RFC4761 - Virtual Private LAN Service (VPLS) Using BGP for Auto :
<http://www.faqs.org/rfcs/rfc4761.html>
- VPLS : Virtual Private LAN Service : Présentation de Jean-Marc Uzé aux JRES 2003 à Lille.
<http://2003.ires.org/actes/paper.134.pdf>
- Documentation JunOS : <http://www.juniper.net>

X. GLOSSAIRE :

- AFI : Address Family Identifier
- BGP : Border Gateway Protocol
- CE : Customer Edge router
- CSPF : Constrained-Path LSP Computation
- FIB : Forwarding Information Base
- LDP : Label Distribution Protocol
- LSP : Label Switched Path
- MPLS : Multiprotocol Label Switching
- MPLS EXP : MPLS Experimental Bits
- NLRI : Network Layer Reachability Information
- OSPF : Open Shortest Path First
- PE : Provider Edge router
- P router : Provider router
- RIB : Routing Information Base
- RSVP-TE : Resource reSerVation Protocol - Traffic Engineering

- SAFI : Subsequent Address Family Identifier
- STP : Spanning-Tree Protocol
- VLAN : Virtual Local Area Network
- VPLS : Virtual Private LAN Service
- VPN : Virtual Private Network
- VSS : Virtual Switching System