Cas d'usage d'un service L3VPN RENATER à l'échelle d'un institut

Jérôme Berthier
Inria DSI Service Conception d'Infrastructure
Centre de recherche Inria Bordeaux – Sud-Ouest
200 avenue de la Vieille Tour
33405 Talence Cedex

Résumé

La DSI Inria a instruit un projet de déploiement d'une solution de téléphonie IP unifiée.

Dans ce cadre, il était nécessaire d'assurer l'ensemble des échanges liés à cette infrastructure (flux signalisation et voix mais aussi accès utilisateurs et administrateurs au portail téléphonie).

Comment adresser et router ces flux TOIP pour connecter les éléments répartis entre les neuf localisations Inria ?

Ce document aborde la problématique initiale et le choix tant au niveau de l'adressage IP qu'au niveau du mécanisme de routage des flux.

Le projet TOIP a démarré en septembre 2015 avec l'impératif de création d'un neuvième site Inria constituant le centre de recherche de Paris (ouverture décembre 2015). L'étude et premier déploiement de la solution L3VPN a été réalisée dans ce délai court par l'équipe réseau de la DSI Inria.

Mots-clefs

RENATER, TOIP, routage, IP, WAN, BGP, MPLS, VRF, PBR, ospfv2, BFD

1 Problématique initiale

1.1 Nouvelle solution TOIP à déployer

Fin 2014, un projet de refonte et de mutualisation autour d'une solution de téléphonie IP unique a été lancé pour aboutir au démarrage d'un déploiement en septembre 2015.

L'architecture globale initiale de la solution retenue est composée de :

- deux serveurs hyperviseurs hébergeant les applicatifs centraux dans le centre d'hébergement mutualisé,
- une passerelle voix permettant la connexion des lignes opérateur T2 sur chaque site,
- une redondance de gestion d'appel virtualisée sur les hyperviseurs existants de chaque site,
- des terminaux répartis selon la population de chaque centre de recherche

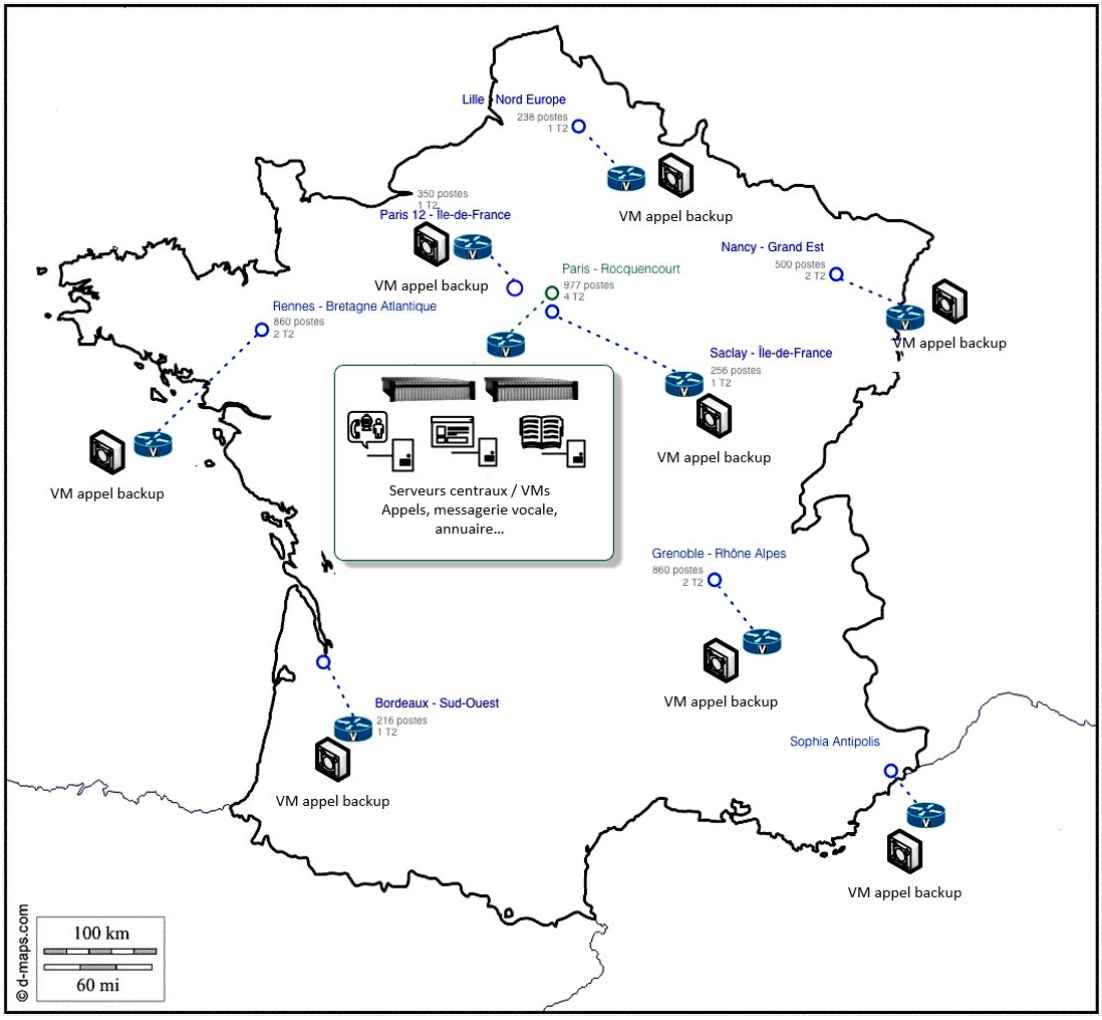

Figure 1: Solution TOIP Inria globale

Cet ensemble constitué de plus de 4500 éléments implique de nombreux échanges de flux transverses au travers du réseau IP :

- flux d'infrastructure : résolution dns, synchronisation ntp, annuaire ldap, …
- flux d'administration des différents éléments (ssh, https, sauvegarde…)
- flux de synchronisation entre serveurs de la solution (gestion des appels, messagerie vocale...)
- flux de paramétrage des terminaux (tftp)
- flux de signalisation TOIP (SIP, SCCP, MGCP)
- flux voix des appels établis (RTP)
- flux de gestion des services utilisateurs (https)

1.2 Problèmes posés par le démarrage de ce projet

La préparation du déploiement de cette nouvelle infrastructure téléphonie IP posa immédiatement deux points à résoudre :

1. Comment adresser l'ensemble de ces éléments à connecter aux réseaux IP Inria ?
2. Comment assurer les échanges entre ces éléments répartis sur les neuf localisations Inria ?

La suite du document aborde les différentes options étudiées, les réponses retenues et leurs modalités de mise en œuvre jusqu'à ce jour.

2 Choix de l'adressage des éléments TOIP

2.1 Trois options d'adressage IP

Chaque site Inria dispose de préfixes IPv4 publics allant de plusieurs classes C à des classes B.

Dans un contexte d'extinction des ressources globales IPv4 disponibles [1], ce choix n'était pas recevable. Il pose par ailleurs des questions de sécurité au niveau des flux UDP notamment.

L'usage de préfixes IPv4 publics a donc été immédiatement écarté.

Chaque site Inria dispose aussi d'un préfixe IPv6 global /48 permettant d'instancier 65536 préfixes /64.

L'usage du protocole IPv6 présentait des incertitudes concernant la compatibilité de la solution TOIP mais aussi de certains équipements réseau Inria à cette époque.

Se baser sur nos préfixes IPv6 globaux était certainement la réponse idéale à notre problématique. Malheureusement, tenant compte des incertitudes de fonctionnement et du planning contraint, cette option n'a pas été retenue.

Notre choix s'est donc porté vers les préfixes IPv4 privés décrits dans le document IETF RFC1918 [2].

2.2 Plan d'adressage RFC1918

Un premier impératif évident était d'éviter tout recouvrement de préfixes entre sites y compris tenant compte des cas d'usage existants.

Par ailleurs, pour offrir de l'évolutivité, ce référentiel devait pouvoir couvrir des besoins autres que le déploiement de la nouvelle solution TOIP.

2.2.1 Attribution d'un préfixe privé à chaque site

Chaque site Inria est doté d'un préfixe /16 dont le second octet vaut le numéro du département de sa localisation géographique.

Le format 10.<département>.0.0/16 permet de traiter chaque site Inria sans recouvrement.

Il y a au moins deux cas d'évolution qui seraient en conflit avec ce modèle :

- ouverture d'un site dans un département d'outre-mer ou à l'étranger
- ouverture d'un second site Inria indépendant d'un premier existant dans le même département

Le cas échéant , la réponse serait alors adaptée au contexte (incrément de 100 ou 200 du numéro de département, référencement de numéros indépendants...).

2.2.2 Attribution d'un préfixe projet TOIP par site

Les éléments de la solution TOIP dépendant d'un site sont adressés dans le préfixe 10.<département>.0.0/21[1].

La localisation principale du site (campus) utilise le préfixe 10.<département>.0.0/22.

Les extensions de localisation (sites distants) utilisent un préfixe sous alloué à partir du préfixe 10.<département>.4.0/22. L'allocation de longueur variable est réalisée selon le besoin.

1. Pour le site d'hébergement principal centralisé, le numéro « département » utilisé est choisi sans lien avec le département réel.

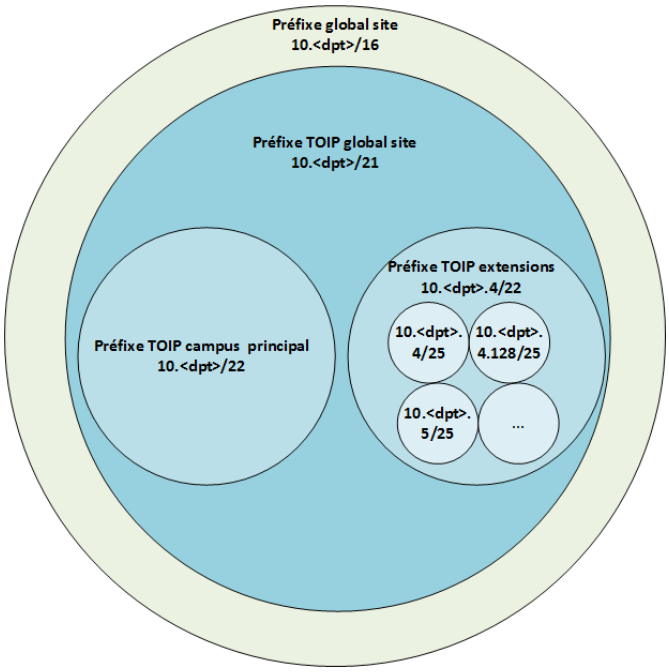

Figure 2: Résumé plan adressage TOIP Inria RFC1918

2.2.3 Problématique de connectivité entre éléments TOIP

Contrairement à un préfixe IPv4 public, un préfixe IPv4 privé n'est pas globalement unique.

Au sein du réseau RENATER, il n'existe pas de règle d'allocation de préfixes privés RFC1918 différenciés par site client.

En conséquence, le trafic concernant un préfixe RFC1918 n'est pas routable nativement sur RENATER (et au delà sur les réseaux constituant des AS[2] différents).

Afin de satisfaire la connectivité IP directe entre les services TOIP de chaque site Inria (incluant le site d'hébergement principal), il était donc nécessaire d'implémenter une solution spécifique.

3 Choix du mécanisme d'échange de trafic TOIP

Plusieurs options ont été évaluées pour assurer la connectivité IPv4 entre les éléments de la solution.

Certains mécanismes sont dédiées au flux de type VOIP, d'autres sont applicables au transport du trafic IP sur le réseau.

Il existe très certainement des variantes ou solutions autres qui ne sont pas abordées ici.

3.1 Usage d'un mécanisme NAT source et destination

Il est possible sur RENATER d'établir une connectivité entre deux hôtes utilisant un adressage privé en empilant deux applications de fonction NAT[3].

Le problème de la traduction d'adresses est la rupture du modèle de connectivité IP de bout en bout.

Or, les protocoles de signalisation TOIP (SIP par exemple) manipulent directement l'adresse du client dans les requêtes d'établissement de sessions.

L'utilisation de traduction NAT rend donc inopérant la signalisation d'appels TOIP et bloque l'établissement du flux voix (RTP).

2. AS = Autonomous System. APNIC, « Autonomous System numbers – FAQs », https://www.apnic.net/get-ip/faqs/asn/

3. NAT = Network Adress Translation. Mécanisme de traduction d'adresses IP sur le réseau

Pour contourner ce problème, il est nécessaire d'activer des fonctionnalités serveur complémentaires afin de permettre la conversation entre hôtes soumis au NAT :

- activation d'un proxy SIP et d'un proxy RTP adressé publiquement et ciblé par les éléments TOIP
- ou activation de serveurs STUN [4] et TURN [5] pour aider les hôtes à connaître et contourner leur application de traduction NAT
- ou activation d'inspection protocolaire avancée sur les pare-feux afin de ré-écrire le contenu tenant compte des sessions NAT

Les mécanismes ci-dessus ont tous pour conséquence de déclencher une modification du trafic émis par les hôtes, parfois à plusieurs niveaux.

Il en relève une complexité accrue de la solution, notamment pour l'analyse de dysfonctionnement.

Afin de garder une cible la plus simple possible, l'utilisation de traduction NAT a été écartée.

Il a été préféré identifier un mécanisme permettant l'échange direct de trafic IP entre les hôtes distants.

3.2 Établissement de trafic IP direct entre sites

Une pratique courante pour rendre deux réseaux distants joignables entre eux est d'établir un tunnel (GRE, IPSEC…) afin d'encapsuler le trafic entre les deux sites et le transporter sur un réseau backbone commun (RENATER ou autre).

L'usage de tunnels nécessite un paramétrage important sur l'ensemble des routeurs WAN concernés pour assurer un maillage complet.

Le transport via un tunnel ajoute aussi des entêtes au trafic IP initial ce qui peut impliquer une fragmentation du trafic en entrée du tunnel.

Afin d'éviter cette fragmentation, il est possible d'agir sur les paramétrages MTU des hôtes, de s'appuyer sur l'algorithme « PMTU discovery » [10] ou finalement de modifier la valeur TCP MSS [11] des sessions.

L'usage de fonctionnalités tunnel notamment IPSEC peut aussi avoir des limitations sur certains équipements (traitement logiciel et non matériel, licence complémentaire…).

Engager un déploiement à base de tunnels impliquait donc une incertitude sur la compatibilité de certains équipements Inria et amenait un impact sur le trafic transporté.

Il a été préféré s'engager vers une solution respectant une connectivité IP de bout en bout sans modification du trafic et s'appuyant uniquement sur des fonctionnalités de routage standard.

Le service L3VPN offert par RENATER a donc été évalué et retenu.

4 Description des services VPN RENATER

4.1 Généralités

Les services VPN font partie de l'offre RENATER [12] permettant d'interconnecter de manière directe différents sites client d'un même établissement ou non.

Les échanges de type communication (voix, vidéo…) sont un exemple d'usage classique mais ce type de service répond aussi à des problématiques de synchronisation entre centres d'hébergement distants, expérimentations réparties…

La solution offerte peut être de niveau 2 (Ethernet - L2VPN) ou de niveau 3 (IP – L3VPN).

Cette offre L3VPN retenue permet d'échanger du trafic IP entre sites distants de manière cloisonnée sur RENATER (voir Annexe 1). Une encapsulation MPLS entre le NR du site source et le NR du site destination assure un transport totalement séparé et sûr des flux concernés sur le backbone.

Les coûts d'utilisation sont à étudier avec RENATER selon l'établissement concerné. De manière classique pour ce type de service, il existe un coût de raccordement (FAS[4]) associé à une redevance annuelle.

4.2 Conditions de raccordement

4.2.1 Demande d'activation du service

Une demande d'adhésion au service doit être adressée au GIP RENATER [14].

Un formulaire de demande de raccordement technique permet de décrire la forme attendu pour le service : sites à raccorder, attributs attendus (débit, métriques de performance, MTU…)...

Il convient d'y préciser le maximum d'éléments notamment : la nécessité de cibler des équipements ou cartes différenciés entre NR et réseau de collecte si possible, le souhait de se connecter à deux NR différents si applicable, l'activation de protocoles spécifiques (exemple BFD)...

4.2.2 Client RENATER direct ou indirect

Il faut différencier le cas d'un site client direct RENATER et d'un site client d'un réseau intermédiaire (exemple : réseau universitaire, métropolitain, régional…).

Un client RENATER direct utilise ses liens physiques existants avec le NR pour s'y raccorder via un nouveau vlan[5] qui portera les échanges L3VPN.

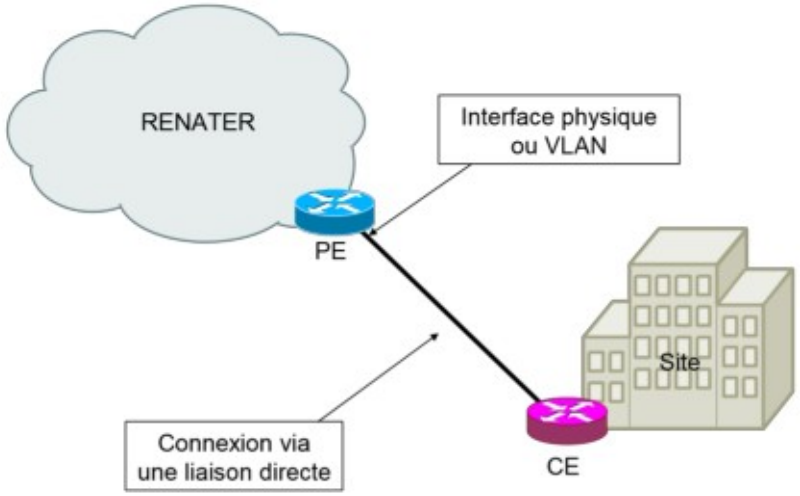

Figure 3: Connexion L2/L3 RENATER directe

(source RENATER – documentation du service L3VPN)

S'il n'existe aucun raccordement direct entre les équipements client et le NR, la liaison sera propagée à travers le réseau intermédiaire via un transport de vlan ou une encapsulation VPLS ou EoMPLS par exemple. Aucun trafic IP L3VPN n'est opéré au niveau du réseau de collecte.

4. FAS = Frais d'accès au service

5. La possibilité d'usage d'une nouvelle interface physique dédiée sur le NR est abordée aussi dans la documentation RENATER.

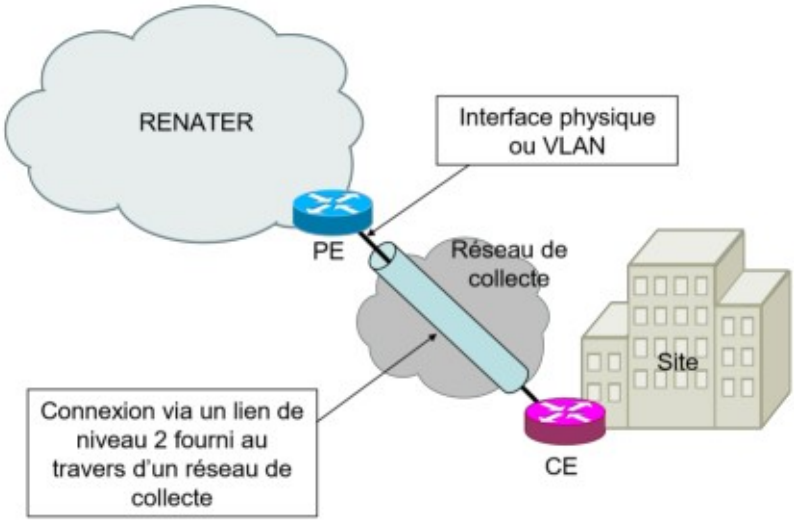

Figure 4: Connexion L2 via Réseau de Collecte et L3 RENATER directe

(source RENATER – documentation du service L3VPN)

La différenciation « client RENATER direct » versus « client réseau collecte » est particulièrement importante. Dans le cas de transit via un réseau de collecte, c'est au client de lui communiquer la demande et la cible technique proposée par RENATER (NR et interfaces).

La disponibilité et le choix du vlan d'interconnexion sont donc dépendants de chaque réseau de collecte et de chaque NR RENATER concernés.

4.2.3 Mise en production RENATER

Après acceptation du raccordement, le GIP RENATER propose une cible technique de mise en œuvre par site (NR, interface, vlan…).

Lorsque tous les attributs sont validés et définis entre toutes les parties (NR concernés, réseau de collecte, interfaces, vlans…) , le GIP RENATER transmet la demande de réalisation au NOC pour mise en production.

En retour, le NOC prend contact avec les correspondants techniques du site pour transmettre les attributs de raccordement : vlans retenus, ASN BGP attendus, adresses IP peering, mot de passe session BGP...

Charge au client de coordonner le déploiement avec les réseaux de collecte pour faire abouter les vlans définis entre les routeurs utilisateurs et le NR RENATER cible.

5 Implémentation L3VPN Inria

5.1 Contexte WAN Inria

5.1.1 Présentation globale

Depuis 2011, la DSI Inria normalise sur chaque site une solution WAN fournissant une haute disponibilité.

Elle est basée sur deux équipements de routage dédiés raccordés à deux points de raccordement amont différents.

Par « deux équipements de routage », il faut considérer deux routeurs séparés ou un seul routeur muni de redondance de composants (doublement des cartes).

Par « deux points de raccordement différents », on entend deux nœuds de raccordement (NR) amont différents ou un seul NR sur lequel on se connecte sur deux cartes différentes.

Chaque équipement de routage Inria est attaché par deux liens physiques[6] fournissant un agrégat LACP vers les coeurs de réseaux.

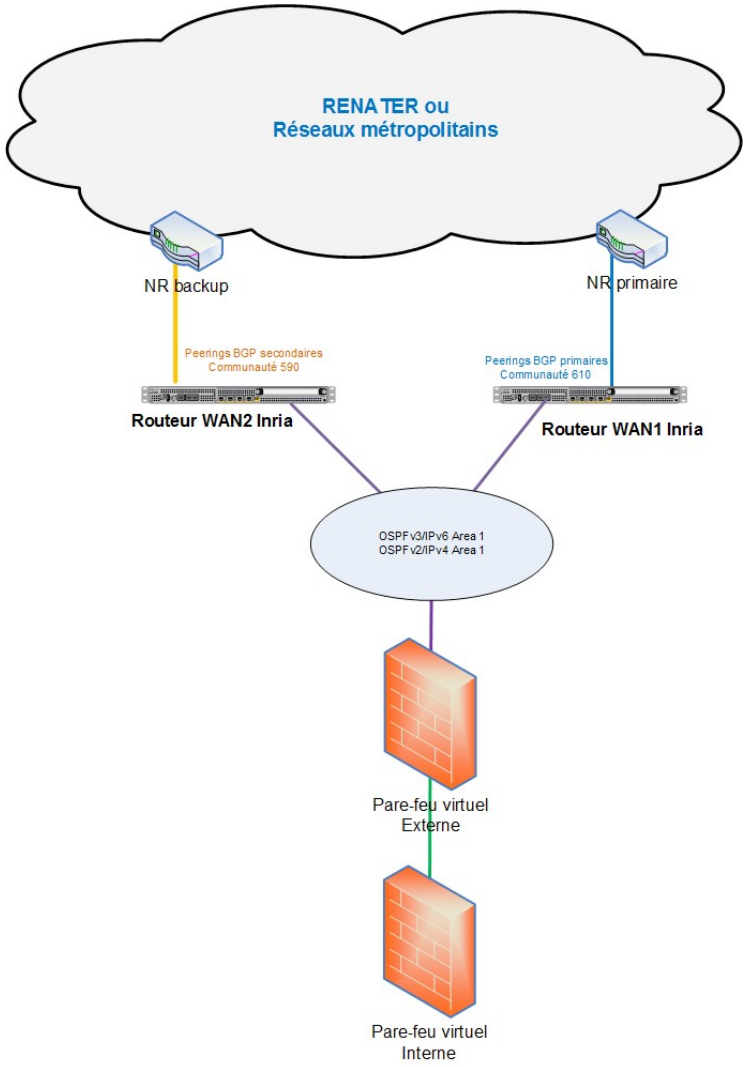

Figure 5: Architecture réseau IP site Inria niveau 3

5.1.2 Routage externe

Le routage IP avec l'extérieur est assuré par deux peerings BGP en IPv4 et en IPv6 sur chaque routeur.

La résilience entre les deux accès repose sur le marquage de communautés [15] appliquées aux annonces BGP[7].

Nos préfixes sont donc annoncés aux réseaux amont de manière identique mais pondérée afin de créer un chemin primaire et secondaire.

5.1.3 Routage interne de site

6. Cette forte redondance physique et l'unicité de l'AS BGP Renater amont nous a permis d'exclure l'usage d'un peering iBGP entre nos routeurs.

7. Dans certains cas, l'usage d'allongement de chemins AS BGP (« AS prepend ») est utilisé pour créer des sous-chemins entre quatre peerings différents. Cela permet d'activer un hébergement simultané sur plus de deux accès différents.

Chaque routeur WAN du site est interconnecté à une instance pare-feu externe via une aire de routage dynamique ospfv2 pour IPv4 et ospfv3 pour IPv6.

Le pare-feu externe est lui même relié au pare-feu interne via du routage statique double pile IPv4 / IPv6.

5.1.4 Cible RENATER pour le L3VPN Inria

Les typologies WAN Inria par site sont de trois types :

- double attachements RENATER direct
- un attachement RENATER + un attachement métropolitain (REAUMUR, ROYAL)
- double attachements métropolitains (TIGRE, Lothaire/StanNet)

Le nouveau vlan d'interconnexion L3VPN avec RENATER est ajouté à l'interface physique qui sert au raccordement publique existant avec RENATER ou réseau métropolitain amont.

Dans notre contexte, la coordination avec les équipes de gestion des différents réseaux amont a été primordiale.

Grace à leur réactivité, nous avons pu rapidement fiabiliser la cible proposée par RENATER, converger vers deux numéros de vlans identiques pour tous les sites et activer la mise en œuvre sans difficulté.

Un même vlan ne pouvant être utilisé deux fois sur le même NR, nous avons introduit une troisième référence pour l'un des NR sur lequel trois terminaisons L3VPN sont réalisées.

5.2 Déploiement L3VPN et fonctionnalités activées

Les configurations proposées sont basées sur des équipements WAN Cisco (voir Annexe 2).

5.2.1 Modification WAN BGP

Chaque routeur WAN Inria porte un nouveau peering BGP L3VPN vers son routeur NR RENATER.

Le système AS RENATER est le numéro 2200.

Dans le cas où le site possède des peerings RENATER publics directs, le numéro AS Inria utilisé est identique.

Dans le cas où le site ne possède pas de peering RENATER public direct, un numéro AS privé lui est affecté par RENATER. Le système AS Inria initial est masqué artificiellement dans les annonces émises. Inversement, le système AS privé affecté par RENATER est supprimé des annonces reçues via les peerings L3VPN.

```
router bgp <ASN_Inria_existant>

neighbor <NR_RENATER_L3VPN_IP_paire> local-as
<ASN_Inria_prive_L3VPN> no-prepend replace-as
```

5.2.2 Politique d'export de préfixes émis par le site

La construction des tables de routage Inria et VRF RENATER vérifie une politique spécifique à l'institut.

Cela permet de maitriser l'horizon de routage visible pour chaque préfixe sur les différents sites.

Les cas de communication identifiés dans le projet TOIP sont les suivants :

- chaque réseau TOIP RFC1918 de site doit voir l'ensemble des autres réseaux TOIP RFC1918
- chaque réseau TOIP RFC1918 de site doit voir les réseaux IPv4 publiques d'infrastructure ou de services du site d'hébergement principal (accès aux ressources centralisées)

– chaque réseau IPv4 public dédié aux utilisateurs de site doit voir le réseau TOIP RFC1918 du site d'hébergement principal (accès aux outils TOIP centralisés)

En terme de routage, cela se traduit par ces cas d'annonces spécifiques :

– Le préfixe RFC1918 de site ne doit jamais être annoncé sur le peering RENATER ou réseau de collecte publique. *Un filtrage strict des annonces IPv4 publiques sortantes doit être effectué.*

– Le préfixe RFC1918 du site doit toujours être annoncé sur le peering RENATER L3VPN.

– Les préfixes IPv4 publics du site doivent aussi être annoncés sur le peering RENATER L3VPN.

Pour trier les annonces, nous utilisons deux mécanismes disponibles au niveau des déclarations de pairs BGP : « distribute-list / ip access list » et «prefix-list / ip prefix-list » (voir Annexe 3).

La priorisation des flux entrants entre l'accès L3VPN primaire et secondaire est réalisée via le marquage de communauté spécifique sur les annonces de préfixes : 610 pour le chemin primaire et 590 pour le chemin secondaire (voir Annexe 4).

Les communautés sont à manipuler dans le format au standard RFC1997 [17]. Si nécessaire, il faut activer ce format :

```
ip bgp-community new-format
```

5.2.3 Politique d'import de préfixes reçus par le site

La politique de routage spécifique à l'institut se décline aussi en entrée de chaque site par la sélection des préfixes intégrés localement à la table de routage globale.

Les préfixes RFC1918 sont toujours être acceptés depuis le peering RENATER L3VPN (voir Annexe 5).

A l'inverse, par défaut, aucun autre préfixe IPv4 public n'est accepté en entrée des sites au risque d'absorber le trafic à destination du centre d'hébergement principal ou d'un autre site via ce service L3VPN[8].

5.2.4 Routage sortant du préfixe RFC1918

Après avoir activé les peerings BGP avec les politiques décrites précédemment, nous avons à ce stade établi un routage effectif entre les plages RFC1918 des différents sites.

Toutefois, les échanges entre préfixes RFC1918 et préfixes IPv4 publics notamment du centre d'hébergement principal ne sont pas encore assurés. Par défaut, le trafic à destination d'un préfixe public sort via l'accès RENATER standard et non par le L3VPN.

8. La bascule vers un transport L3VPN pour les autres flux entre préfixes Inria publics est planifiée au cas par cas.

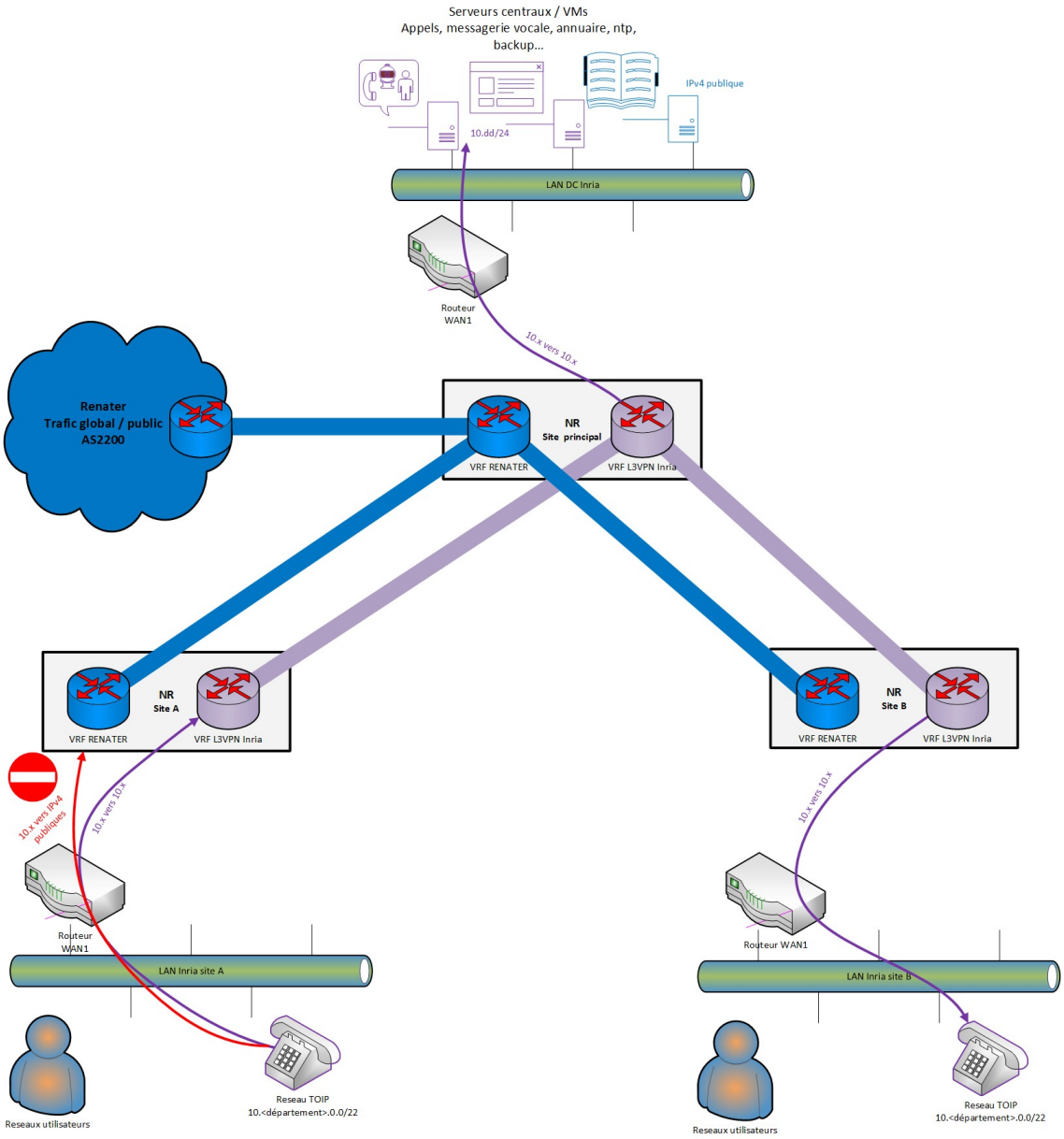

Figure 6: Flux simplifiés avant PBR (un seul NR et routeur Inria)

Le préfixe RFC1918 de site devrait *toujours* sortir via le peering BGP L3VPN y compris à destination des plages IPv4 publiques qui y sont annoncées.

Il faut donc influer spécifiquement sur le routage sortant du préfixe RFC1918 de site pour l'aiguiller[9].

Cette action est réalisée via une fonction de routage par la source (PBR = Policy Based Routing) appliquée sur l'interface interne de chaque routeur WAN Inria (voir Annexe 6).

Cette politique PBR décrit une logique spécifique :

« Pour tout paquet entrant sur l'interface interne dont l'adresse source appartient au préfixe RFC1918 de site et dont l'adresse destination n'appartient pas au site alors l'interface de sortie et le next hop sont le peering BGP L3VPN. »

9. L'usage de vrf locales et redistribution aurait pu apporter un cloisonnement plus flexible (vrf lite).

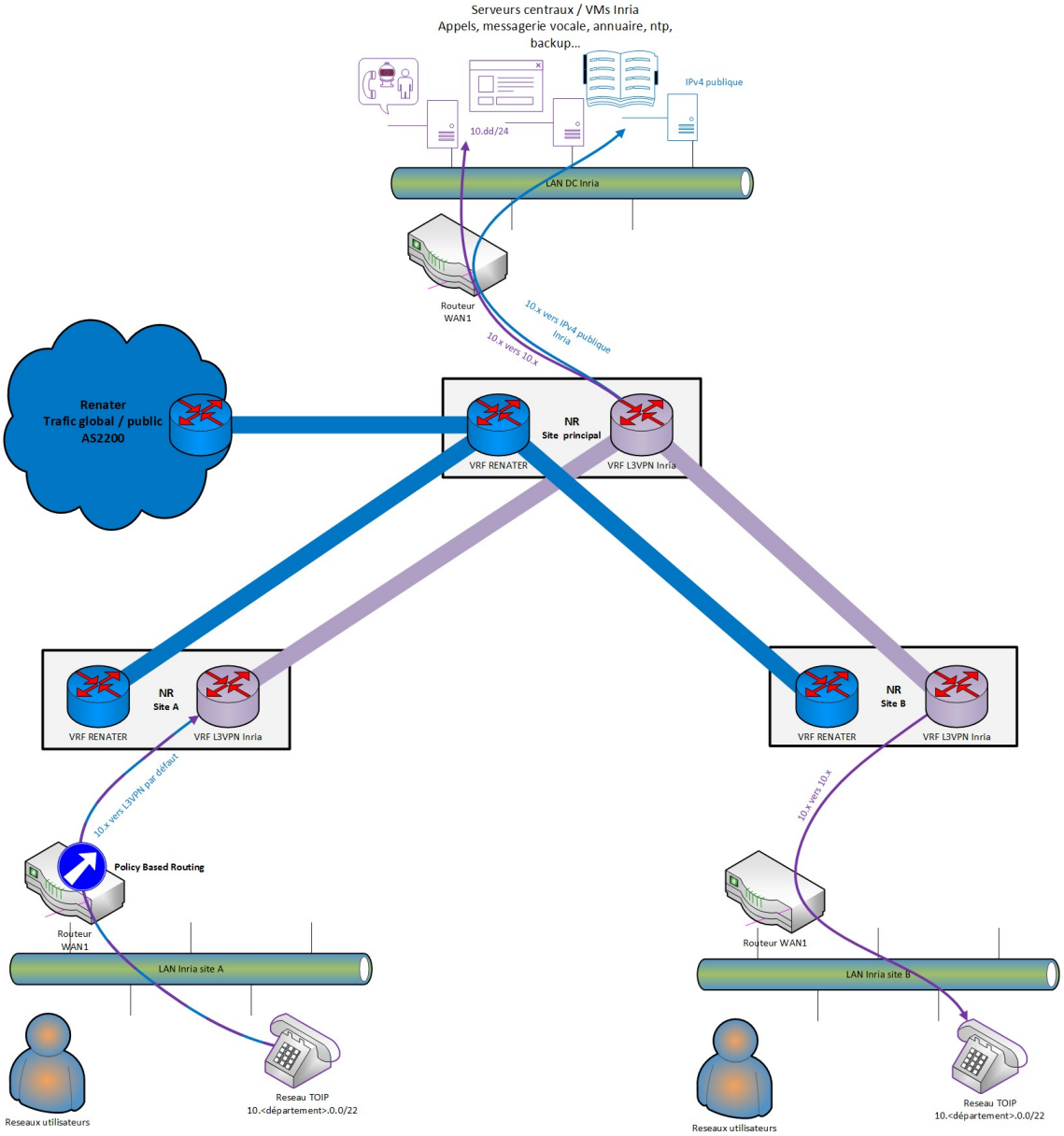

Figure 7: Flux simplifiés avec PBR (un seul NR et routeur Inria)

5.2.5 Détection de panne et convergence rapide BGP

Afin d'accélérer la convergence du trafic en cas de perte d'un accès, le protocole BFD est utilisé pour tester la présence du voisin BGP et déclencher la fin de session et purge des tables de routage en cas de dysfonctionnement (voir Annexe 7).

Ce test de vie est particulièrement utile sur les transports de niveau 2 indirects entre le site et le NR (transit via un réseau tiers). Un transit niveau 2 implique des cas de panne sans perte de l'interface physique utilisée par le peering qui reste établi mais non fonctionnel (timer hold 180s par défaut).

Le timing de test BFD validé avec RENATER est très agressif : hello toutes les 150ms et 5 échecs avant de considérer le voisin comme hors service. Il faut donc uniquement 750ms pour détecter une panne.

Ces délais très courts sont très sensibles à toute dégradation de trafic (perte de paquets) au risque de se déclencher anormalement. C'est une limitation à prendre en compte.

5.2.6 Filtrage IP « sanitaire »

En entrée du L3VPN, un filtrage par access-list sans état est appliqué[10].

Cette liste de filtrage bloque :

- les sources réseaux du site pour protéger contre l'usurpation d'adresses IP source (anti-spoofing)
- les préfixes « anormaux » sur un peering BGP
- les messages icmp non souhaités
- les connexions directes sur le routeur WAN Inria à l'exception des flux BGP et BFD

In fine, cette liste de filtrage laisse passer tout trafic à destination des réseaux du site.

En sortie vers le L3VPN, aucun filtre n'est appliqué.

L'introduction de préfixes RFC1918 sur le site implique une vigilance supplémentaire sur le trafic susceptible de sortir par les peering RENATER ou métropolitain publics.

Afin d'éviter toute fuite de trafic issu ou à destination d'adresses IP privées, il est préférable d'appliquer un filtrage en sortie des peerings existants[11]. Cela fait d'ailleurs partie des préconisations référencées dans le document BCP38 [18] afin d'améliorer la lutte contre les dénis de service à base d'usurpation d'adresses IP source.

Cette liste de filtrage bloque :

- les réseaux RFC 1918 en source ET destination
- les préfixes « anormaux »
- tout trafic non issu d'un préfixe IPv4 du site

Tout filtrage plus spécifique est ensuite à la charge des équipements pare-feu du site.

6 Bilan L3VPN et évolutions

Généralisé en 2016 à tous les sites Inria, l'usage du service L3VPN RENATER est un véritable atout pour accompagner la mise en œuvre de solutions transverses à l'institut.

Dès le début, nous avons profité de ce service pour activer aussi un mécanisme de redondance auprès de notre opérateur téléphonie fixe Bouygues Telecom.

En complément de la solution TOIP, cela permet dynamiquement de faire déborder et sécuriser les appels entrants en cas de saturation ou panne de l'accès ISDN T2 d'un site. Ces appels sont alors reroutés par l'opérateur sur un autre site Inria et renvoyé via le L3VPN vers les éléments TOIP du site en panne.

Depuis ce déploiement initial, nous avons mis en œuvre deux autres cas d'usage L3VPN :

- une solution d'impression centralisée adressée aussi via les préfixes RFC1918
- une redistribution de flux SSL VPN utilisateurs entre les concentrateurs et les ressources IPv4 publiques des différents sites

Les évolutions d'usage possible de ce service L3VPN sont nombreuses : réflexion sur le traitement d'autres flux internes Inria (partiel ou généralisation) mais surtout *déploiement d'un service L3VPN IPv6.*

Les débits mesurés sur les peerings L3VPN restent faibles de l'ordre de quelques mégabits[12]. L'injection d'autres flux Inria transverses fera rapidement monter le trafic mesuré.

A l'image des autres services RENATER, les pannes sont quasi inexistantes.

10. Un filtre équivalent est déjà appliqué au peering RENATER standard.

11. Cette préconisation reste à généraliser sur tous nos peerings.

12. Mesures hors flux VPN SSL utilisateurs en cours de mise en production

C'est un service fiable et performant qui n'a aucun coût d'exploitation interne hors déploiement. Dans le cadre de projets, nous opérons uniquement du paramétrage pour adapter les politiques d'import / export BGP et PBR afin d'accompagner les nouveaux usages internes.

Un point de vigilance demeure concernant l'activation de QOS WAN afin de privilégier les flux critiques entre le vlan IP publique et le vlan L3VPN. La saturation globale de l'interface reste un risque non traité à ce stade (pas uniquement critique pour la TOIP).

Annexe

Annexe 1 - fonctionnement technique L3VPN résumé

L'offre L3VPN permet d'échanger du trafic IP entre sites distants de manière cloisonnée sur RENATER.

Du côté utilisateur, le service L3VPN RENATER est basé sur l'établissement d'un peering eBGP dédié entre le NR RENATER (équipement PE MPLS) et le routeur des sites clients à raccorder (équipement CE).

Le cloisonnement du trafic inter-site est assuré par le protocole MPLS qui transporte les flux sans modification du trafic IP initial.

L'encapsulation est réalisée par l'ajout de labels "entre" la couche 2 (Ethernet) et la couche 3 (IP) du modèle OSI. Plusieurs labels sont manipulés : les labels de service identifiant le VPN et les labels de commutation identifiant le chemin backbone entre les NR concernés (déterminé via la signalisation LDP ou RSVP).

Cette encapsulation MPLS entre le NR du site source et le NR du site destination assure un transport totalement séparé et sûr des flux concernés sur le backbone. Aucun chiffrement de trafic n'est appliqué.

L'établissement des peerings eBGP (via la famille d'adresses « vpn4 » côté NR) est associé à la création d'une table de routage IP virtuelle (VRF) locale et dédiée sur chaque NR RENATER concernés.

Côté site client, une fois son peering BGP établi, chaque routeur de site envoie ses informations sur les préfixes IP (les routes) qu'il souhaite faire connaitre.

Le NR RENATER pair en alimente sa table de routage virtuelle associée au service L3VPN du client.

Les préfixes de cette table de routage virtuelle sont exportés et partagés via le protocole MP-BGP famille « vpn4 » ((Network Layer Reachability Information (NLRI) - type VPN-IPv4) afin que les autres NR concernés par ce même service L3VPN (attribut Route Target (RT) commun) puissent intégrer les informations de routage à leur propre table VRF locale.

Chaque NR RENATER transmet alors aux routeurs de site concernés l'ensemble des préfixes IPv4 qu'il a intégré dans sa table VRF locale (ie les préfixes injectés par les autres routeurs CE dans ce service L3VPN).

Les sites connectés au L3VPN sont entièrement libres des informations de routage qu'ils envoient aux pairs BGP RENATER. La seule limite est quantitative à hauteur de 200 préfixes pour l'ensemble de l'instance L3VPN.

La méthode de fourniture de services VPN backbone BGP / MPLS est décrite dans le standard RFC 4364 [13].

Annexe 2 - socle matériel et logiciel

Les configurations proposées sont basées sur des équipements WAN Cisco famille ASR1000 exécutant le système IOS 15 (équivalent IOS-XE 3.10 et supérieur)[13].

La fonctionnalité Cisco Express Forwarding (CEF) [16] doit être activée pour IPv4 (état de base sur les systèmes cités).

Les mécanismes utilisés sont standardisés donc a priori utilisables sur d'autres solutions propriétaire ou libre (exemple projets FRRouting (FRR)[14] ou VyOS[15]).

Annexe 3 – filtrage de préfixes émis par le site

Exemple peering public – filtre « bgp-filter-v4-out » - blocage par défaut incluant les annonces RFC1918

13. Il ne devrait pas y avoir d'incompatibilité particulière pour appliquer cette solution à des équipements plus anciens sous IOS 12.

14. Projet FRRouting (FRR) ; https://frrouting.org/

15. Projet VyOS ; https://vyos.io/

```
ip prefix-list bgp-filter-v4-out seq 5 permit
<préfixe1_public>/32

ip prefix-list bgp-filter-v4-out seq 10 permit <préfixe2_public>/
18

ip prefix-list bgp-filter-v4-out seq 15 permit <préfixe3_public>/
17

ip prefix-list bgp-filter-v4-out seq 20 permit <préfixe4_public>/
24

ip prefix-list bgp-filter-v4-out seq 25 permit <préfixe5_public>/
24 le 32

ip prefix-list bgp-filter-v4-out seq 30 permit <préfixe6_public>/
28

ip prefix-list bgp-filter-v4-out seq 35 permit <préfixe7_public>/
29 le 32

ip prefix-list bgp-filter-v4-out seq 100 deny 0.0.0.0/0 le 32
```

Exemple peering L3VPN – filtre «prefix-L3VPN-v4-out » - annonces des préfixes retenus

```
ip access-list standard prefix-L3VPN-v4-out

 permit <préfixe1_public_site> <masque inversé ACL Cisco>

 permit <préfixe2_public_site> <masque inversé ACL Cisco>

 permit 10.<département>.0.0 0.0.255.255

 deny   any
```

Ces filtrages d'annonces sont appliqués vers le pair BGP RENATER en sortie :

```
router bgp <ASN>
```

```
 address-family ipv4

  neighbor <IP_pair_RENATER_public> prefix-list bgp-filter-v4-out
out

  neighbor <IP_pair_RENATER_L3VPN> distribute-list prefix-L3VPN-
v4-out out
```

Annexe 4 – application communautés BGP

Une access-list « BGP-COMMUNITY-to-L3VPN-v4 » est utilisée pour sélectionner les réseaux candidats à l'application de la communauté.

Cette liste contient le réseau RFC1918 du site et l'ensemble des réseaux IPv4 publics du site (en fait un contenu identique à la liste «prefix-L3VPN-v4-out » précédente).

Selon le routeur WAN[16], la variable <communauté> sera donc fixée à 610 ou 590[17] via une liste d'actions route-map « SET-BGP-COMMUNITY-L3VPN-v4 ».

```
route-map SET-BGP-COMMUNITY-L3VPN-v4 permit 10

 match ip address BGP-COMMUNITY-to-L3VPN-v4

 set community 2200:<communauté>

!

route-map SET-BGP-COMMUNITY-L3VPN-v4 permit 20
```

Ce traitement est ensuite appliquée au pair BGP RENATER en sortie :

```
router bgp <ASN>

 address-family ipv4

  neighbor <IP_pair_RENATER_L3VPN> route-map SET-BGP-COMMUNITY-
L3VPN-v4 out
```

Annexe 5 - filtrage de préfixes reçus par le site

16. Si un seul routeur porte les deux peerings primaire et secondaire, la configuration varie pour traiter les deux cas.

17. L'absence de marquage de communauté 590 n'empêcherait pas la priorisation car par défaut, RENATER applique une valeur « local pref » à 600 lors des imports donc valeur inférieure à 610 passée sur l'accès primaire.

Pour cela en entrée du L3VPN, une access-list « prefix-L3VPN-v4-in » est créée et contient uniquement le réseau RFC1918 10/8 commun.

```
ip access-list standard prefix-L3VPN-v4-in

 permit 10.0.0.0 0.255.255.255

 deny   any
```

Elle est ensuite appliquée sous forme de « distribute-list » entrante associée au pair BGP L3VPN sur chaque routeur WAN Inria.

```
router bgp <ASN>

 address-family ipv4

    neighbor <IP_pair_RENATER_L3VPN> distribute-list prefix-
L3VPN-v4-in in
```

Exemple peering L3VPN – routes RFC1918 intégrées des préfixes reçus

```
      10.0.0.0/8 is variably subnetted, 16 subnets, 5 masks

B        10.6.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d

B        10.33.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d

B        10.35.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d

B        10.38.0.0/16 [20/0] via <IP_pair_RENATER>, 3d06h

B        10.54.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d

B        10.59.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d

B        10.78.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d

B        10.91.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d
```

```
B        10.101.0.0/16 [20/0] via <IP_pair_RENATER>, 1w1d
```

Annexe 6 – application PBR

Une access-list « ACL-map-RFC1918-to-IPv4-global » est créée pour sélectionner les réseaux candidats à l'application du routage PBR. Cette liste contient le réseau RFC1918 du site et exclut les préfixes publics du site.

```
ip access-list extended ACL-map-RFC1918-to-IPv4-global

deny ip 10.<département>.0.0 0.0.255.255 <prefix1_site_Inria>
<wildcard_mask>

deny ip 10.<département>.0.0 0.0.255.255 <prefix2_site_Inria>
<wildcard_mask>

!...

permit ip 10.<département>.0.0 0.0.255.255 any

deny ip any any
```

Cette access-list est ensuite appliquée via une liste d'actions route-map « RT-traffic-to-L3VPN » :

```
route-map RT-traffic-to-L3VPN permit 10

match ip address ACL-map-RFC1918-to-IPv4-global

set interface <interface_physique>.<vlan>

set ip next-hop <RENATER_L3VPN_IP_paire>



route-map RT-traffic-to-L3VPN permit 20
```

Si les deux peering L3VPN sont portés sur un unique routeur, la forme de la route-map doit être adaptée afin d'utiliser une hiérarchie dans les next hop sans assignation de l'interface de sortie. La directive « set interface null0 » y sera ajoutée pour détruire le trafic en cas d'indisponibilité des deux peers L3VPN.

```
! route-map si routeur Inria unique

route-map RT-traffic-to-L3VPN permit 10

match ip address ACL-map-RFC1918-to-IPv4-global

set ip next-hop <RENATER_Peer1_L3VPN_IP_paire>

set ip next-hop recursive <RENATER_Peer2_L3VPN_IP_paire>

set interface null0



route-map RT-traffic-to-L3VPN permit 20
```

L'application du routage PBR est directement réalisé en entrée de l'interface « interne » de chaque routeur. Cette interface correspond à celle utilisée pour faire la liaison avec le pare-feu externe du site via l'aire ospf. Il s'agit d'une interface marquée 802.1Q sur un agrégat LACP Port-channel1.

```
interface Port-channel1.<vlan_interne>

ip policy route-map RT-traffic-to-L3VPN
```

Annexe 7 – application BFD

Le mécanisme de test est déclaré au niveau de l'interface IP d'interconnexion vers le NR RENATER (et réciproquement) :

```
interface <interface_physique>.<vlan>

description L3VPN

encapsulation dot1Q <vlan>

ip address <Inria_L3VPN_IP_impaire> 255.255.255.254

bfd interval 150 min_rx 150 multiplier 5
```

```
no bfd echo

![…]
```

Le protocole BFD est ensuite associé au voisin BGP RENATER. Ainsi, tout changement d'état BFD provoque la fermeture immédiate de la session BGP avec le voisin concerné :

```
router bgp <ASN>

neighbor <IP_pair_RENATER_L3VPN> fall-over bfd
```

Le mode echo permet normalement à l'équipement de s'auto-adresser le paquet hello en fixant sa propre adresse IP en source *ET destination* puis en transmettant le paquet à son voisin. Cela crée une situation d'aller retour direct permettant d'alléger le traitement du protocole qui est restreint alors au niveau du plan de traitement (data / forwarding plane) et non au niveau du plan de contrôle (processeur central).

Toutefois, ce mode ne fonctionne pas avec RENATER qui semble appliquer un filtrage ou vérification de chemin inverse[18] interdisant ce trafic.

Annexe 8 – filtrage IP sanitaire

Exemple - filtre «de-L3VPN » - filtrage du trafic reçu

```
ip access-list extended de-L3VPN

 remark anti-spoofing plages <site>

 deny   ip <prefix1_Inria_site> <wildcard_mask> any log

 deny   ip <prefix2_Inria_site> <wildcard_mask> any log

 deny   ip 10.<département>.0.0 0.0.255.255 any log

 remark filtrage classes RFC1918 hors classe 10.0.0.0/8

 deny ip 172.16.0.0 0.15.255.255 any log

 deny ip 192.168.0.0 0.0.255.255 any log
```

18. « Unicast Reverse Path Forwarding ». Cisco, https://www.cisco.com/c/en/us/about/security-center/unicast-reverse-path-forwarding.html

```
remark filtrage classes RFC3927

deny   ip 169.254.0.0 0.0.255.255 any log

remark filtrage classes RFC1122

deny   ip 127.0.0.0 0.255.255.255 any log

remark filtrage classes RFC1700

deny   ip 0.0.0.0 0.255.255.255 any log

remark filtrage classes RFC5737

deny   ip 192.0.2.0 0.0.0.255 any log

deny   ip 198.51.100.0 0.0.0.255 any log

deny   ip 203.0.113.0 0.0.0.255 any log

remark ICMP

permit icmp any any echo

permit icmp any any echo-reply

permit icmp any any net-unreachable

permit icmp any any reassembly-timeout

permit icmp any any time-exceeded

permit icmp any any ttl-exceeded

permit icmp any any unreachable

permit icmp any any administratively-prohibited

deny   icmp any any log

remark flux bgp peer RENATER
```

```
 permit tcp host <RENATER_L3VPN_IP_paire> host
<Inria_L3VPN_IP_impaire> eq bgp

 remark retour BFD echo src dst identique

 permit udp host <Inria_L3VPN_IP_impaire> host
<Inria_L3VPN_IP_impaire> range 3784 3785

 permit udp host <RENATER_L3VPN_IP_paire> host
<RENATER_L3VPN_IP_paire> range 3784 3785

  remark INTERDIT SUR LES ROUTEURS

 deny   ip any host <Inria_interfaces_routeur> log

 deny   ip any host <Inria_L3VPN_IP_impaire> log

 remark tout IP L3VPN vers prefixes du site

 permit ip any 10.<département>.0.0 0.0.255.255

 permit ip any <prefix1_Inria_site> <wildcard_mask>

 permit ip any <prefix2_Inria_site> <wildcard_mask>

 remark fermeture finale

 deny   ip any any log
```

Exemple - filtre «vers-INTERNET-BCP38 » - filtrage du trafic public sortant

```
ip access-list extended vers-INTERNET-BCP38

 remark deny vers classes RFC1918

 deny   ip any 10.0.0.0 0.255.255.255 log

 deny   ip any 172.16.0.0 0.15.255.255 log

 deny   ip any 192.168.0.0 0.0.255.255 log
```

```
 remark deny vers classes RFC3927

 deny   ip any 169.254.0.0 0.0.255.255 log

 remark deny vers classes RFC1122

 deny   ip any 127.0.0.0 0.255.255.255 log

 remark deny vers classes RFC1700

 deny   ip any 0.0.0.0 0.255.255.255 log

 remark deny vers classes RFC5737

 deny   ip any 192.0.2.0 0.0.0.255 log

 deny   ip any 198.51.100.0 0.0.0.255 log

 deny   ip any 203.0.113.0 0.0.0.255 log

 remark sources prefixes publiques <site>

 permit ip <prefix1_Inria_site> <wildcard_mask> any

 permit ip <prefix2_Inria_site> <wildcard_mask> any

 remark permit depuis routeurs interface peering

 permit ip host <Inria_L3VPN_IP_impaire> host
<RENATER_L3VPN_IP_paire> eq bgp

 permit udp host <RENATER_L3VPN_IP_paire> host
<RENATER_L3VPN_IP_paire> eq 3784

 permit udp host <RENATER_L3VPN_IP_paire> host
<RENATER_L3VPN_IP_paire> eq 3785

 remark fermeture finale

 deny   ip any any log
```

Annexe 9 – commandes opérationnelles utiles

– Voisinage BGP

```
show ip bgp summary
```

– Voisinage BFD

```
show bfd neighbors
```

– Préfixes BGP reçus d'un peering

```
show ip bgp neighbors <IP_neighbor> received-routes
```

– Préfixes BGP émis sur un peering

```
show ip bgp neighbors <IP_neighbor> advertised-routes
```

– Préfixes BGP intégrés à la table de routage globale

```
show ip route bgp
```

Annexe 10 – supervision

La supervision du service est réalisée via le protocole SNMPv3 AuthPriv pour vérifier l'état de chaque peering.

Le serveur doit disposer de la MIB CISCO-BGP4-MIB :

– lien : ftp://ftp.cisco.com/pub/mibs/v2/CISCO-BGP4-MIB.my

– à copier sous /usr/share/snmp/mibs/ sous CentOS 7

L'état de chaque peering BGP est récupérable via une requête snmpget sur oid 1.3.6.1.4.1.9.9.187.1.2.5.1.3.1.4.<IP_peer> pour les peerings IPv4.

Un peering BGP peut vérifier les états suivants :

1 -> Idle

2 -> Connect

3 -> Active

4 -> OpenSent

5 -> OpenConfirm

6 -> Established

Chaque peering doit être établi donc renvoyer la valeur numérique 6 lors de l'interrogation snmp.

Bibliographie

[1] Article « IPv4 exhaustion ». Wikipedia, https://en.wikipedia.org/wiki/IPv4_address_exhaustion

[2] Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. de Groot, E. Lear. Document « Address Allocation for Private Internets ». IETF, Février 1996 ; https://tools.ietf.org/html/rfc1918

[3] Article « IPv4 exhaustion ». Wikipedia, https://en.wikipedia.org/wiki/IPv4_address_exhaustion

[4] Article « Session Traversal Utilities for NAT (STUN) ». Wikipedia, https://en.wikipedia.org/wiki/STUN

[5] Article « Traversal Using Relays around NAT (TURN) ». Wikipedia, https://en.wikipedia.org/wiki/Traversal_Using_Relays_around_NAT

[6] Article « Generic Routing Encapsulation (GRE) ». Wikipedia, https://en.wikipedia.org/wiki/Generic_Routing_Encapsulation

[7] Article « Internet Protocol Security (IPSEC) ». Wikipedia, https://fr.wikipedia.org/wiki/IPsec

[8] Document « Cisco IOS DMVPN Overview ». Cisco, Février 2008; https://www.cisco.com/c/dam/en/us/products/collateral/security/dynamic-multipoint-vpn-dmvpn/DMVPN_Overview.pdf

[9] Documentation « Juniper Auto Discovery VPNs ». Juniper, 09 Juillet 2019; https://www.juniper.net/documentation/en_US/junos/topics/topic-map/security-auto-discovery-vpns.html

[10] Article « Path MTU discovery (PMTUd) ». Wikipedia, https://fr.wikipedia.org/wiki/Path_MTU_discovery

[11] Ivan Pepelnjak. Article « TCP MSS clamping – what is it and why do we need it? ». Blog ipSpace, 22 Janvier 2013; https://blog.ipspace.net/2013/01/tcp-mss-clamping-what-is-it-and-why-do.html

[12] Documentation « Services VPN ». RENATER; https://www.renater.fr/fr/Services%20VPN

[13] E. Rosen, Y. Rekhter. Document « BGP/MPLS IP Virtual Private Networks (VPNs) ». IETF, Février 2006; https://tools.ietf.org/html/rfc4364

[14] Documentation « Demande de VPN ». RENATER; https://services.renater.fr/vpn/index

[15] Documentation « Communautés BGP ». RENATER; https://services.renater.fr/services_ip/routage_d_adresses#communautes_bgp

[16] Documentation « IP Switching: Cisco Express Forwarding Configuration Guide, Cisco IOS Release 15S ». Cisco, 24 Janvier 2018; https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/ipswitch_cef/configuration/15-s/isw-cef-15-s-book/isw-cef-enable-disable.html?referring_site=RE&pos=1&page=https://www.cisco.com/c/en/us/support/docs/ios-nx-os-software/ios-software-releases-120-mainline/47205-cef-whichpath.html

[17] R. Chandra, P. Traina, T. Li. Document « BGP Communities Attribute ». IETF, Août 1996; https://tools.ietf.org/html/rfc1997

[18] P. Ferguson, D. Senie. Document « Network Ingress Filtering:Defeating Denial of Service Attacks which employ IP Source Address Spoofing ». IETF, Mai 2000; https://tools.ietf.org/html/bcp38