Faciliter l'usage de services réseaux avancés : le cas de SIRES 2 sur RAP

Catherine Grenet, Ludovic Ishiomin

Faciliter l'usage de services réseaux avancés : le cas de SIRES 2 sur RAP

Catherine Grenet
Unité Réseaux du CNRS

Ludovic Ishiomin
Centre Opérationnel de RAP

Résumé

Fiabilisation des raccordements, VPN de niveau 2 ou de niveau 3 : l'accès simultané à ces services avancés est désormais possible, à la condition que ceux-ci ne complexifient pas l'architecture à déployer.

RAP est le réseau de la communauté enseignement supérieur et recherche de Paris. SIRES est le réseau privé virtuel dédié au système d'information de gestion du CNRS.

La présentation détaillera l'évolution de l'architecture de routage sur RAP en insistant sur la démarche d'ingénierie adoptée. Celle-ci a eu pour résultat de proposer des mécanismes simples qui peuvent être généralisés au niveau des sites. Puis nous montrerons comment, dans ce cadre, la conception de l'architecture et la mise en place du réseau SIRES 2 sur les sites raccordés à RAP se sont trouvées grandement simplifiées.

En effet, le réseau SIRES vient d'être migré vers le service de réseau privé virtuel de niveau 3 (VPN MPLS/BGP) proposé par RENATER. Nous montrerons comment dans le cas des six sites connectés à RAP, le service offert par RENATER peut être prolongé jusqu'aux sites, et comment l'architecture mise en place sur RAP peut permettre à ces sites de bénéficier de services avancés tels que le raccordement fiabilisé au VPN de manière transparente.

Mots clefs

L3VPN, BGP, VPN MPLS/BGP, SIRES, ingénierie.

1 Le Réseau Académique Parisien

Le Réseau Académique Parisien est le réseau de la communauté enseignement supérieur et recherche de Paris. Il connecte entre eux et avec RENATER les 140 sites des 65 établissements présents sur le réseau.

Le projet RAP a démarré en 1998, et le réseau est opérationnel depuis 2002 après une première étape de déploiement. Il a connu plusieurs phases d'évolution, que ce soit au niveau du nombre de sites raccordés, ou bien au niveau des services offerts. Au cours de l'année 2008, l'ensemble des actifs du réseau a été complètement renouvelé. Cette opération a eu pour conséquence de considérablement simplifier l'architecture de routage, ce qui a permis d'offrir aux utilisateurs du réseau des services de niveau 3 avancés.

1.1 Architecture du routage à l'origine

Le réseau RAP est bâti autour d'un anneau optique de 5 points de présence géographiquement distribués dans Paris éclairé en Gigabit. Chaque site est raccordé à son point de présence le plus proche par une liaison optique à 100 Mégas. Cela donne une topologie de 5 étoiles reliées par un anneau.

Les sites sont raccordés à un équipement de concentration et de commutation de niveau 2 et 3. Le point de présence de Jussieu est hébergé dans les mêmes locaux que le nœud RENATER de Jussieu.

RAP a été conçu au départ comme un réseau de campus étendu. Par souci de simplification, et pour répondre à un objectif de migration rapide des sites de leur précédent accès au réseau vers RAP, le routage statique entre le site et RAP a été choisi. Un IGP[1] est mis en place sur l'anneau, et transporte les réseaux d'interconnexion ainsi que ceux des sites. Le choix s'est porté sur

[1] *Interior Gateway Protocol*

OSPF[2]. Ce fut d'ailleurs plus un choix « par défaut » (il fallait un IGP, donc pourquoi pas celui-là !) qu'un choix étayé objectivement.

Un équipement de bordure était chargé d'opérer la connexion avec RENATER. Il était connecté à l'équipement de concentration de Jussieu. Il récupérait les routes OSPF et les annonçait en BGP[3] à RENATER. Il générait en retour une route par défaut dans l'OSPF.

Cette architecture, représentée à la figure 1, possède le mérite de la simplicité. Elle ne présente pas de difficultés particulières concernant l'exploitation quotidienne du réseau, ce qui est avantageux lorsqu'on est en phase active de déploiement. L'ajout d'un nouveau site est une opération dont l'impact est facilement maîtrisé.

Sur RAP, certains sites dits « multi-homés » disposent d'une liaison vers un fournisseur extérieur. Ils possèdent un numéro d'AS public leur permettant d'avoir deux accès Internet indépendants. Un *peering* BGP *multi-hop*[4] entre leur équipement et celui de bordure avec RENATER leur permet de récupérer la table de routage globale de l'Internet, en plus des routes RAP. On remarque déjà une faiblesse dans la capacité à converger rapidement selon l'endroit où se situe la panne pour un site « multi-homé » : à son niveau, la convergence peut être rapide en cas de panne sur le lien de son deuxième accès ; à cause du « *multihop* », elle nécessite d'attendre la temporisation BGP en cas de problème sur le chemin emprunté pour atteindre l'équipement de bordure RAP.

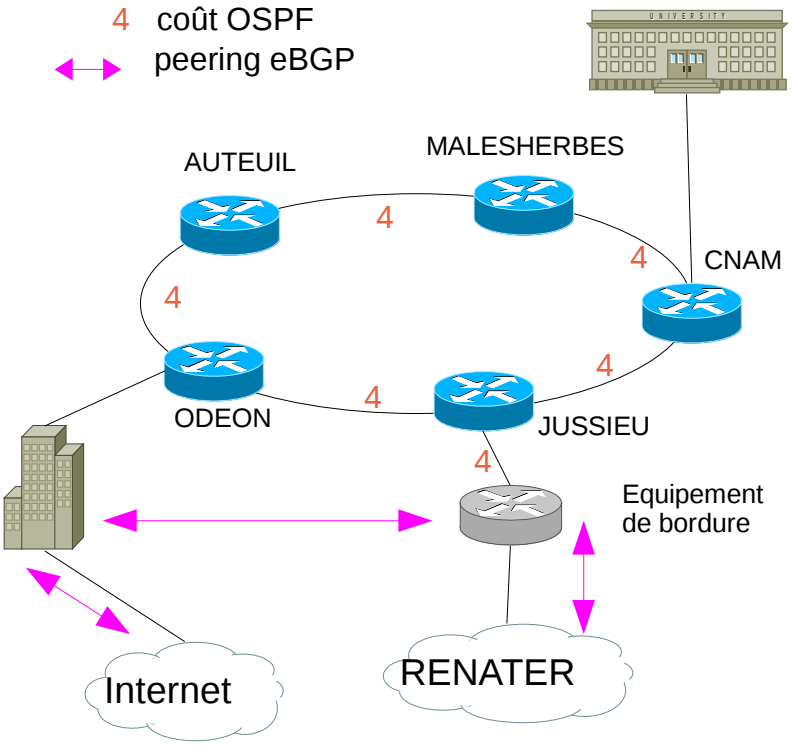

Figure 1 – *Architecture de routage à l'origine de RAP*

1.2 Mise en œuvre de BGP pour les sites

Le déploiement plus conséquent de BGP sur le réseau est survenu fin 2003 lorsqu'un établissement a émis le souhait de bénéficier d'un mécanisme permettant à deux de ses sites de se secourir mutuellement, sachant qu'il disposait d'une liaison privée entre ses deux sites. Au Centre Opérationnel de RAP, nous avons estimé que d'autres établissements seraient susceptibles d'adopter une solution similaire pour certains sites dits sensibles. Nous avions donc l'objectif de concevoir une solution qui ne soit pas spécifique à cet établissement demandeur. La contrainte fixée consistait à ne pas modifier le mode de raccordement des autres sites en production.

Le protocole de choix, pour échanger dynamiquement ne serait-ce que quelques routes entre domaines différents, est BGP. Un IGP nous paraissait exclu principalement à cause de l'absence de contrôle des annonces de routes reçues du site. De plus, BGP permet de mettre en place assez facilement une ingénierie de routage en jouant simplement sur différents critères. Par

[2]*Open Shortest Path First*

[3]*Border Gateway Protocol*

[4]Par opposition à une session BGP entre deux routeurs adjacents.

conséquent, nous avons déployé une infrastructure iBGP[5] entre nos équipements. Nous avons mis en place un *route-reflector*[6] sur l'équipement de bordure, afin de simplifier les configurations et d'éviter de déclarer au minimum 20 *peerings* sur chaque équipement. La figure suivante schématise l'ajout de BGP sur RAP. La destination des routes BGP est résolue récursivement sur chaque équipement via OSPF.

Toutefois, après cette étape de déploiement, une évolution de la topologie de routage s'est avérée nécessaire. En effet, de part l'augmentation continue du trafic, du fait du grand nombre de sites raccordés, et pour des raisons liées à leur architecture matérielle interne, les équipements des points de présence de Jussieu et Odéon n'étaient pas capables de traiter correctement au niveau 3 le trafic en transit depuis l'extérieur vers Auteuil, Malesherbes et Cnam. Pour pallier ce problème, une étoile entre l'équipement de bordure avec RENATER et chacun des cinq points de présence a été construite à base de VLAN de niveau 2, en plus de l'anneau. La figure 2 représente la topologie logique de routage résultante.

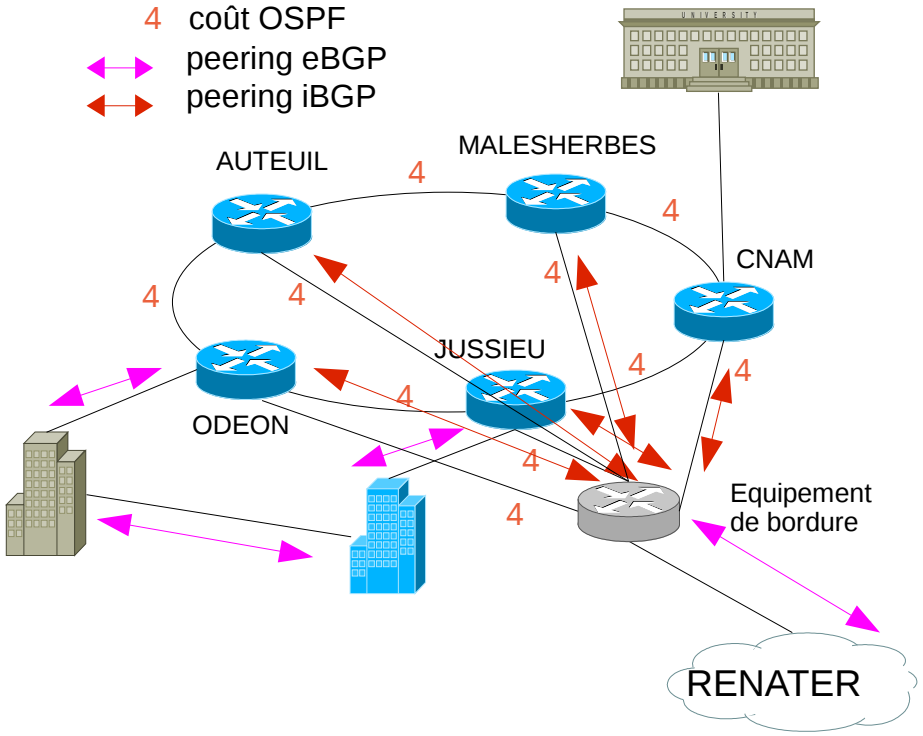

Figure 2 – *Topologie logique de routage après déploiement de BGP et évolution liée à l'augmentation de trafic*

1.3 Le second raccordement à RENATER

Le deuxième raccordement à RENATER, intervenu fin 2006, a été l'opportunité pour RAP d'améliorer significativement la disponibilité globale du réseau pour les utilisateurs. L'objectif visé consistait à disposer d'une redondance d'accès à RENATER, en mettant en place une deuxième liaison depuis un autre point de présence que Jussieu, vers un NR[7] différent. Un deuxième équipement de bordure était prévu pour ce projet. Les contraintes que nous nous sommes posées étaient les suivantes :

- cette deuxième liaison ne devait pas seulement servir en cas de panne ou de maintenance de la première : le meilleur moyen de s'assurer qu'elle fonctionne correctement, c'est de l'utiliser en permanence ;
- la nouvelle architecture ne doit pas modifier du point de vue des sites leur mode de raccordement à RAP ;
- nous souhaitions éviter en temps normal le trafic asymétrique sur le réseau.

Le mécanisme offert par RENATER pour annoncer une même route à deux NR différents consiste à lui ajouter une communauté BGP différente sur chaque NR. RENATER traduit alors ces communautés en une priorité : il est ainsi possible d'indiquer à RENATER par quel NR la route devient prioritairement accessible. D'autres mécanismes dans BGP permettant de rendre prioritaire une route par rapport à une autre sont possibles (par exemple l'attribut MED[8]). L'avantage de celui-ci est qu'il rend explicite l'ingénierie de trafic dans la configuration des équipements.

[5] *Internal BGP*

[6] Le comportement par défaut d'un routeur BGP consiste à ne pas annoncer aux autres voisins BGP du même AS (*Autonomous System*) les routes apprises d'un voisin ayant le même AS ; le *route-reflector* centralise toutes ces routes pour les redistribuer à tous les autres voisins appartenant au même AS.

[7] *Nœud RENATER*

[8] *Multi Exit Discriminator*

Le réseau a donc été partitionné en deux : d'un côté les sites raccordés aux points de présence de Jussieu et Cnam voient leur trafic transiter en temps normal par le NR de Jussieu. De l'autre, le trafic des sites d'Odéon, Auteuil et Malesherbes emprunte le deuxième raccordement aboutissant au NR d'Aubervilliers (dénommé Paris1 depuis RENATER 5). Ce découpage n'est bien sûr pas fait par hasard. Nous avions identifié que le trafic global entre RAP et RENATER était à peu près équitablement réparti entre ces deux zones.

Par conséquent, la communauté envoyée à RENATER découle des coûts OSPF des routes des sites, de manière à privilégier l'entrée dans RAP au plus près des sites. Autrement dit, il s'agit de minimiser le nombre d'équipements RAP à traverser compte-tenu du partitionnement décrit précédemment.

Pour les sites « multi-homés », un deuxième *peering* eBGP multihop vers le deuxième équipement de bordure a été mis en place. Ainsi, ces sites bénéficient aussi de la redondance d'accès vers RENATER même en cas de panne ou de maintenance sur l'une des liaisons. La communauté envoyée à RENATER pour les routes de ces sites est définie statiquement dans la configuration des équipements de bordure.

En ce qui concerne les sites qui se secourent mutuellement, la communauté envoyée à RENATER découle de la métrique des routes BGP (attribut « localpref »), qui joue un rôle similaire au coût OSPF décrit ci-avant. Nous leur avons demandé de configurer la route par défaut que nous leur annonçons de manière à ce que chaque site envoie le trafic sortant plutôt via la liaison RAP que via la liaison inter-site, et ce afin d'éviter tout trafic asymétrique.

Le schéma de la figure 3 résume la topologie de routage résultat de l'ingénierie de trafic mise en place.

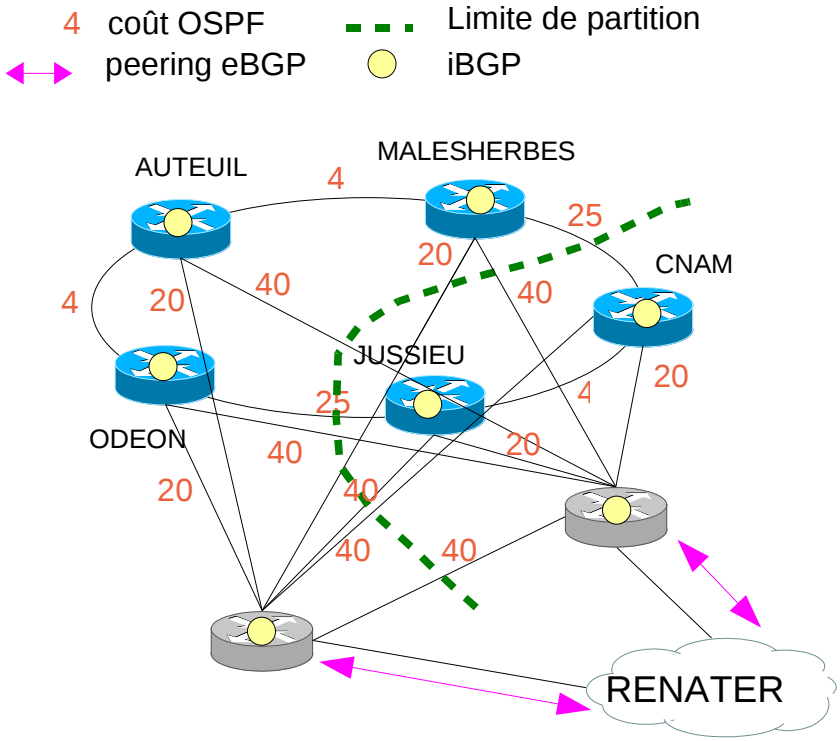

Figure 3 – *Architecture de routage suite au deuxième raccordement RENATER*

1.4 Le raccordement fiabilisé des sites

Un site en raccordement fiabilisé sur RAP dispose de deux liaisons vers deux points de présence différents. La deuxième liaison peut être construite de deux manières : en réutilisant une partie des infrastructures de la première liaison du site (notamment les câbles en égout), ce qui minimise l'investissement nécessaire, ou bien en construisant une infrastructure redondante pour l'accès du site à RAP. Cette capacité à offrir un deuxième accès aux sites a été rendue possible par le renouvellement du marché des liaisons optiques des sites qui a pris effet en 2007.

À cette occasion, nous avons souhaité étendre vers les sites en raccordement fiabilisé notre capacité à mettre en œuvre une répartition des trafics vers RENATER. Le moyen le plus simple consistait à définir cinq communautés BGP [1] [2], qui se traduisent sur chaque point de présence par une métrique (attribut « localpref »), transportée dans l'iBGP RAP[9]. Un site en raccordement fiabilisé doit mettre en place un *peering* eBGP avec RAP sur chaque liaison, en utilisant le même AS. Pour

[9]Ces cinq communautés et leur métrique associée sur chaque point de présence forment un ordre partiel ; un ordre total sera nécessaire dès lors qu'un site aura plus de deux accès à RAP.

chaque route annoncée par le site, selon la communauté envoyée (identique sur les deux *peerings*), son trafic entrant (de RAP vers le site) empruntera en temps normal une liaison (donc un point de présence) plutôt que l'autre. Par ailleurs, nous demandons au site de configurer son trafic sortant de manière à éviter tout trafic asymétrique[10].

La solution adoptée, outre le fait qu'elle rend explicite dans la configuration des équipements RAP le mécanisme utilisé pour le raccordement fiabilisé, permet au site de choisir une des trois stratégies suivantes :

- concentrer les deux liaisons d'accès à RAP sur un seul équipement ; dans ce cas de figure, le plus simple consiste à envoyer une seule communauté pour toutes les routes du site, ce qui revient à n'utiliser qu'une seule liaison en temps normal, l'autre servant en cas de panne de la première ;
- utiliser deux équipements différents pour les deux accès, mais comme précédemment une seule liaison active en temps normal ; le plus simple consiste à mettre en place entre les deux équipements un protocole tel que VRRP, de manière à ce que le réseau interne du site ne « voie » qu'une seule sortie ;
- utiliser deux équipements différents, en mettant en place une ingénierie de trafic dans le réseau du site pour équilibrer la charge des deux liaisons, et potentiellement disposer en temps normal de deux fois le débit nominal d'une seule liaison.

Les annonces des routes des sites en raccordement fiabilisé faites à RENATER découlent alors de la métrique de la route (attribut BGP *local-preference*) déduite de la communauté envoyée par le site.

Afin de rendre homogène les configurations de nos équipements, nous avons demandé aux sites se secourant mutuellement de mettre en place cet envoi de communauté BGP[11].

1.5 Renouvellement des équipements

Un rapide bilan des évolutions décrites précédemment fait apparaître les caractéristiques suivantes :

- environ 650 routes de sites transportées en OSPF, traduites sur les équipements de bordure avec RENATER en communautés BGP en fonction de leur coût ;
- une dizaine de routes de sites transportées en BGP, traduites en communautés BGP vers RENATER en fonction de leur métrique (attribut BGP *local-preference*).
- la modification de la partition entre points de présence pour privilégier telle ou telle sortie sur RENATER (afin, par exemple, de rééquilibrer le trafic RAP-RENATER entre les deux partitions) n'est vraiment pas simple, car elle nécessite de reprendre à la fois l'ensemble des coûts OSPF sur le réseau, ainsi que la traduction de ces coûts en communautés BGP vers RENATER, pour prendre en compte l'ensemble des cas possibles de pannes.

Le lecteur pourra se référer à [3] pour connaître les détails liés au renouvellement des équipements du réseau. Le résultat de cette opération d'envergure a été une simplification drastique de l'architecture de routage. En effet, nous n'utilisons plus que BGP pour les routes des sites, qu'elles soient reprises des routes statiques sur chaque point de présence ou annoncées dynamiquement par les sites. La métrique de ces routes est transportée dans l'iBGP de RAP. OSPF reste en vigueur pour découvrir la topologie et déterminer le plus court chemin en tout point du réseau. La configuration des équipements s'est trouvée grandement simplifiée.

Au final, la métrique BGP d'une route de site ou d'une route provenant de RENATER permet de déterminer le point de présence depuis lequel elle émane. Cette métrique est envoyée aux autres points de présence de manière à implémenter la partition du réseau pour les routes provenant de RENATER, ou bien l'ingénierie de trafic dynamiquement demandée pour les routes des sites en raccordement fiabilisé.

La figure 4 montre la simplification de l'architecture suite à ce renouvellement opéré en 2008.

[10]RAP annonce une route par défaut, à charge pour le site de privilégier le lien utilisé pour son trafic sortant.

[11]Au choix des deux sites qui se secourent mutuellement, un AS unique ou deux AS différents.

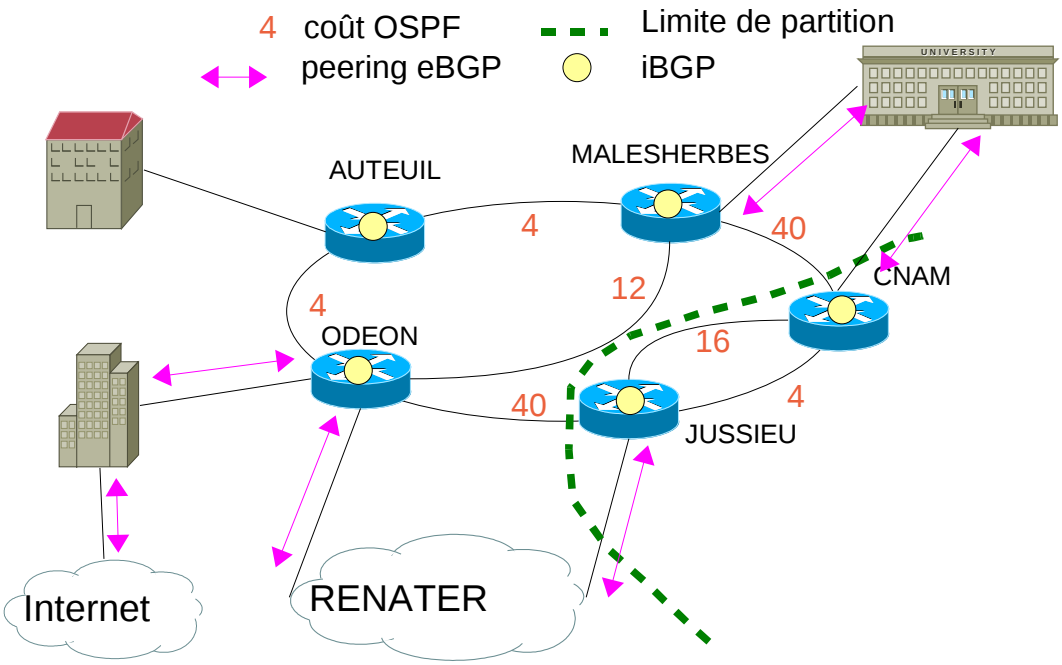

Figure 4 – *Simplification du routage suite au renouvellement des équipements*

2 Le réseau privé virtuel SIRES

2.1 Historique

SIRES est le réseau privé virtuel[12] dédié au système d'information de gestion du CNRS. Il véhicule les flux liés aux applications de gestion, au nombre d'une centaine et dont les deux principales sont la gestion financière et comptable et la gestion des ressources humaines, entre les sites de l'administration du CNRS. L'objectif de SIRES, lors de sa mise en place en 1999, était de garantir aux applications confidentialité, disponibilité et qualité de service.

SIRES interconnecte vingt-trois sites en France : le centre serveur de Trélazé hébergé chez Bull dans la banlieue d'Angers, le centre serveur de Villeurbanne hébergé au centre de calcul de l'IN2P3, les deux sites de la Direction des Systèmes d'Information du CNRS situés respectivement dans les banlieues de Paris et Toulouse, le siège du CNRS à Paris et dix-huit délégations régionales réparties sur le territoire national. L'implantation des différents sites est représentée à la figure 5.

[12]En anglais *Virtual Private Network*, en abrégé VPN : c'est l'acronyme que nous utiliserons par la suite.

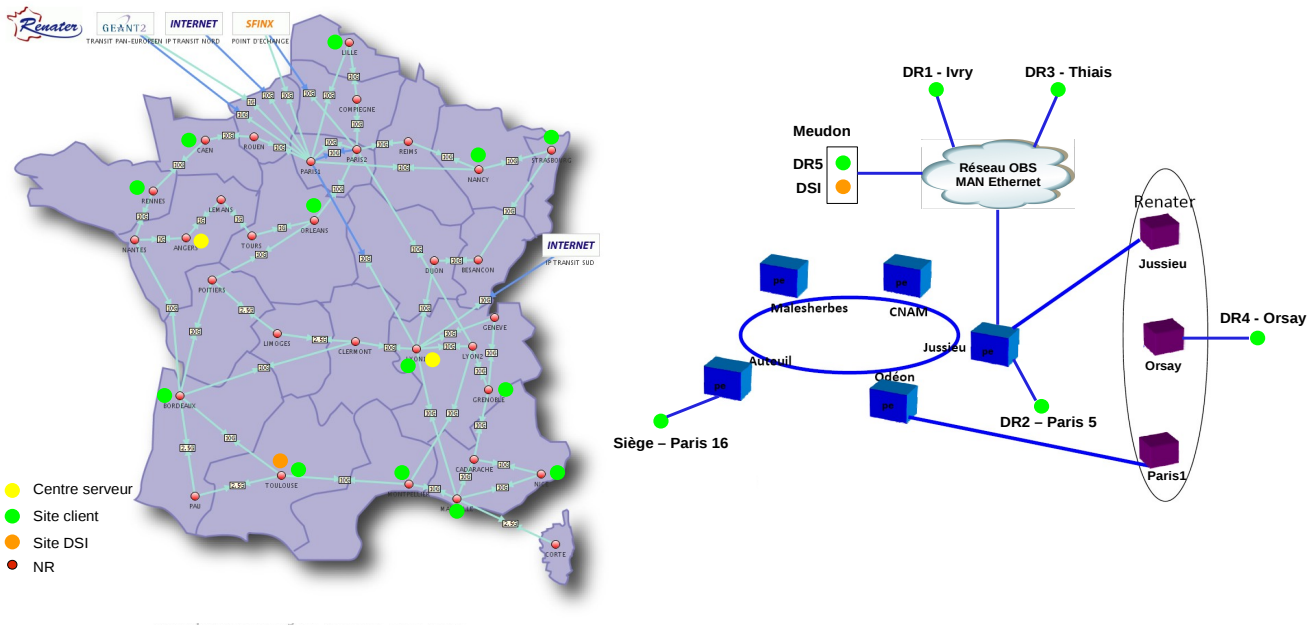

Figure 5 – *Implantation des sites SIRES en France (à gauche) et en Ile-de-France (à droite)*

La première version de SIRES fut construite à l'aide de circuits virtuels ATM[13] sur l'infrastructure de RENATER 2, organisés en étoile autour d'un routeur central. La confidentialité et la qualité de service étaient assurées par les mécanismes intrinsèques à ATM, la disponibilité par des liens RNIS qui doublaient ceux allant de chaque site au routeur central. Chacun des sites était alors relié à un NR par l'intermédiaire d'une liaison louée. Seul le trafic entre les sites SIRES empruntait le réseau SIRES, le reste du trafic était routé normalement sur RENATER.

En 2002, lorsque RENATER abandonna ATM, SIRES migra vers un ensemble de tunnels IP dans IP, GRE[14] ou IPsec, toujours organisés en étoile autour du routeur central. Puis les liaisons spécialisées furent abandonnées partout où c'était possible, au profit des réseaux métropolitains et régionaux. Les liaisons de secours RNIS furent remplacées par des liaisons ADSL. Le protocole de routage utilisé sur le réseau était OSPF.

En 2007, le CNRS mit en service de nouvelles applications de gestion basées sur SAP, qui centralisaient davantage les traitements dans les centres serveurs. Les conséquences pour le réseau furent une augmentation des débits entre le centre serveur de Trélazé et les autres sites et un besoin accru de disponibilité. Pour diminuer la charge du routeur central sur lequel aboutissaient l'ensemble des tunnels GRE, on ne routa plus dans SIRES que le trafic entre le centre serveur de Trélazé et les autres sites. Cet épisode mit en évidence les limitations liées à l'architecture de SIRES :

- le protocole GRE, exécuté par du logiciel dans les routeurs, a un fort impact sur les performances des équipements ;
- l'utilisation des tunnels pose des problèmes spécifiques tels que la fragmentation ;
- enfin l'architecture en étoile n'est pas la mieux adaptée pour un réseau où tous les sites doivent communiquer entre eux.

Nous décidâmes donc de faire évoluer le réseau SIRES vers une nouvelle version appelée SIRES 2.

2.2 Architecture de SIRES 2

La nouvelle version du VPN SIRES devait intégrer la totalité du trafic entre les sites et s'appuyer sur l'infrastructure des réseaux de la recherche, c'est-à-dire RENATER et les réseaux métropolitains et régionaux. Le choix fut fait d'utiliser le service de réseau privé virtuel de niveau 3, ou L3VPN, de RENATER, qui est un service de VPN MPLS/BGP tel que défini par [4].

[13] *Asynchronous Transfert Mode*

[14] *Generic Routing Encapsulation*

L'utilisation de ce service permet de bénéficier d'un réseau entièrement maillé, où il est très facile de rajouter un site. La complexité pour les équipes d'exploitation est réduite puisque celle-ci est gérée par l'opérateur. Le confort est très proche de celui du réseau local, en particulier grâce au gain en performance résultant de la suppression des tunnels GRE. De plus ce service offre, ou devrait offrir prochainement, des possibilités telles que le raccordement en IPv6, le multicast et la qualité de service, que nous n'utilisons pas actuellement car nous n'en avons pas besoin.

Il nous a cependant fallu adapter l'architecture de routage de notre réseau. En effet le protocole de routage utilisé sur SIRES était OSPF. Il aurait été possible de conserver OSPF sur le réseau de secours tout en utilisant BGP sur le réseau nominal, c'est-à-dire le L3VPN, mais cela conduisait à des configurations relativement complexes. C'est pourquoi nous avons préféré migrer l'ensemble du réseau en BGP. Cependant l'utilisation de BGP avec les paramètres par défaut faisait passer le temps de bascule sur la liaison de secours en cas d'incident à quelques minutes. L'utilisation du protocole BFD[15] a permis de réduire ce délai à quelques secondes.

2.3 Mise en œuvre de SIRES 2

Le point d'entrée du VPN est, pour chaque site, le NR de RENATER. Le routeur de chaque site établit une session BGP avec le routeur de RENATER, dans laquelle il annonce ses réseaux et reçoit les annonces des réseaux des autres sites appartenant au VPN. Cette session BGP est établie dans un VLAN dédié, en parallèle d'un second VLAN par lequel transite le trafic non SIRES. La figure 6 illustre sur ce principe pour un site directement raccordé au NR.

Cependant, si certains des sites de SIRES sont directement raccordés à un NR, la majorité d'entre eux est connectée à RENATER par l'intermédiaire d'un réseau métropolitain ou régional, voire les deux. Or l'accès au service de L3VPN de RENATER n'est pas possible par l'interconnexion standard à un réseau régional ou métropolitain. Nous avons donc demandé aux équipes en charge de ces différents réseaux s'il était possible d'établir un lien de niveau 2 entre le site et le NR. Tous nous ont répondu positivement et nous ont proposé d'utiliser pour cela un VLAN 802.1Q, à l'exception d'un qui nous a proposé une fibre optique dédiée. Le schéma de principe est illustré figure 7. Le VLAN « SIRES » transporte la session BGP avec le NR, tandis que dans le VLAN « non SIRES » on garde l'interconnexion habituelle : routage statique, OSPF ou BGP.

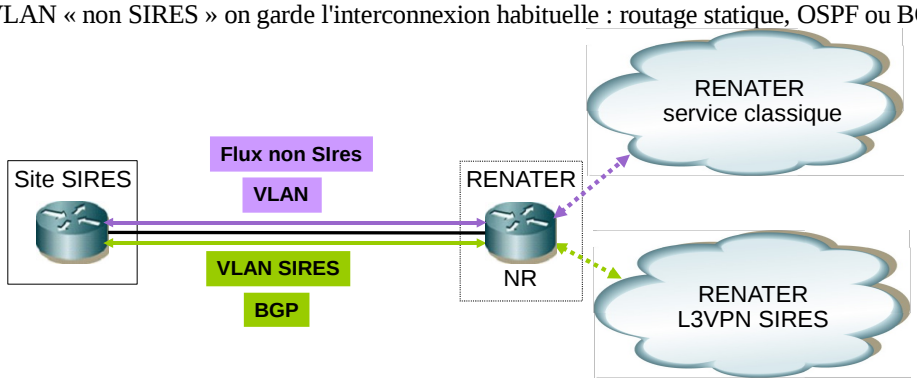

Figure 6 – *Site directement raccordé à un NR*

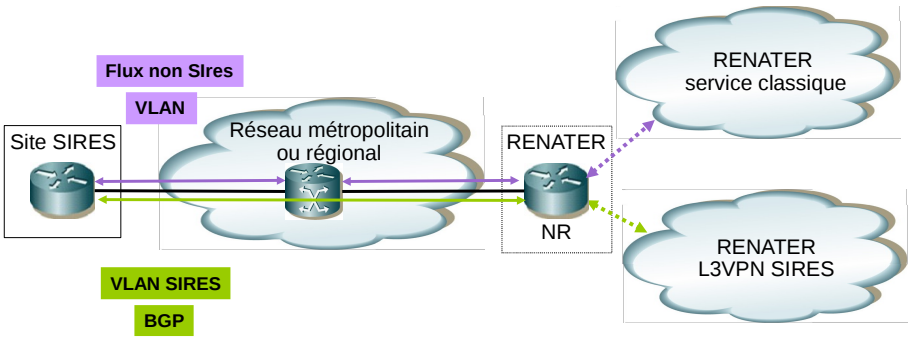

Figure7 – *Site raccordé à travers un réseau métropolitain ou régional*

[15]*Bidirectional Forwarding Detection*. Ce mécanisme repose sur l'échange de messages entre les équipements : lorsqu'un routeur ne reçoit plus de messages il fait tomber la session BGP sans attendre l'expiration des temporisations.

3 SIRES 2 sur RAP

3.1 Premier déploiement

Les six sites de SIRES connectés à RAP ont été raccordés au L3VPN à l'été 2009, suivant le principe décrit au paragraphe 2.3, c'est-à-dire par des L2VPN reliant chacun des sites au NR de Paris1, comme le montre la figure 8.

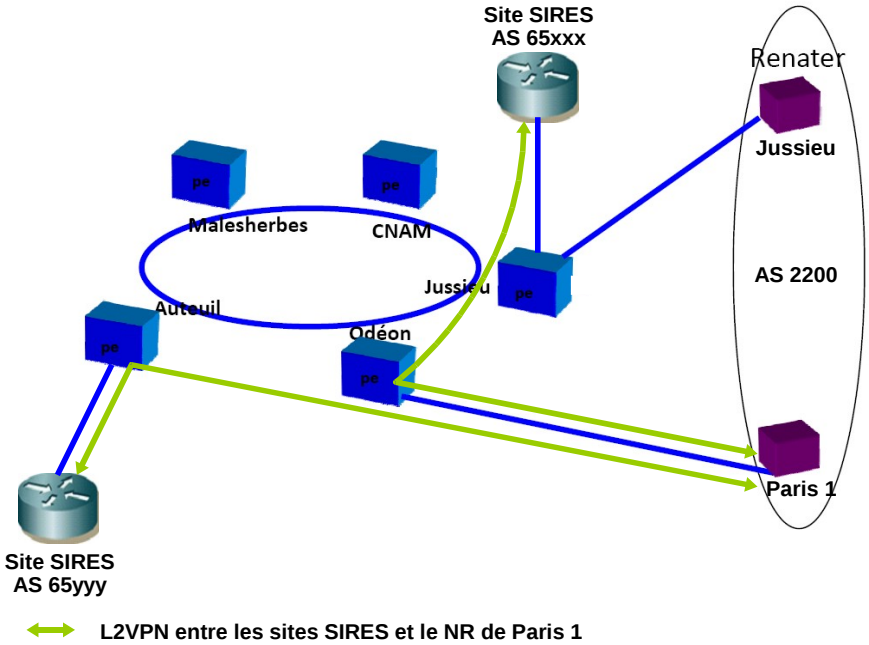

Figure 8 – Premier déploiement sur RAP

Cette architecture de déploiement présente un inconvénient majeur : l'accès de niveau 1 et 2 est rendu par une entité (RAP), le service de niveau 3 par une autre (RENATER). Il en résulte potentiellement une difficulté à solliciter le support en cas d'incident. Idéalement, il faudrait déterminer à quel niveau se situe le problème, de manière à ne faire intervenir que l'acteur concerné. En pratique, cette démarche est très difficile à réaliser. Il est nécessaire de contacter conjointement RAP et RENATER pour un premier diagnostic. Si cela ne résout pas l'incident, alors la procédure devient bien plus complexe, car elle nécessite une multiplicité d'allers-retours entre chaque acteur sur les résultats de diagnostics partiels. Il n'est pas prévu à l'avance qu'un des acteurs prenne le rôle « d'intégrateur » afin de piloter les autres pour résoudre le problème au plus vite[16].

3.2 Les L3VPN sur RAP

L'infrastructure MPLS sur RAP se base sur des LSP[17] définis entre chaque point de présence (et ce dans chaque sens, puisqu'un LSP est unidirectionnel), signalés par RSVP[18]. Ceci constitue le transport des paquets utilisé par les différents services (VPLS[19] et L3VPN). Une description plus précise de cette infrastructure dépasse le cadre de cet article. Notons simplement que celle-ci offre des mécanismes de FRR[20] de l'ordre de la centaine de millisecondes.

Le mécanisme de transport des routes des L3VPN dans l'iBGP est légèrement différent par rapport à celui mis en place pour le service de routage IPv4 ou IPv6. En effet, pour ceux-ci, le *next-hop* est résolu par l'IGP. En MPLS, ce *next-hop* se matérialise par un label MPLS, représentant l'équipement vers lequel le datagramme sera acheminé pour être délivré à l'utilisateur

[16]Pour témoigner de la difficulté à résoudre au plus vite un incident sur un autre L3VPN déployé de cette manière pour le compte d'un autre établissement : le site RAP de celui-ci s'est retrouvé isolé pendant 15 jours de son L3VPN, et ce malgré la volonté évidente des 4 acteurs (l'établissement en question, l'autre réseau métropolitain impliqué dans le L3VPN, RENATER et RAP) à trouver la solution. L'incident était finalement l'accumulation de plusieurs problèmes en différents points.

[17]*Label Switched Path*

[18]*Resource Reservation Protocol*

[19]*Virtual Private Lan Service*

[20]Fast ReRoute

(RENATER ou le site). De plus, chaque route émise par un routeur RAP (un routeur « PE[21] » dans la terminologie MPLS) devient unique dans l'iBGP RAP par ajout du « *route-distinguisher*[22] ».

La répartition des flux du L3VPN vers les deux accès RENATER se fait de la manière suivante. Pour les flux sortants (de RAP vers RENATER) :

- chaque routeur de bordure de RAP apprend les routes du L3VPN annoncées par RENATER, en les marquant d'une communauté particulière et les réannonce dans l'iBGP de RAP ;
- les autres équipements intègrent ces routes du VPN depuis l'iBGP vers leur VRF[23] en traduisant cette communauté en métrique, de telle manière que ceux d'Auteuil et Malesherbes privilégient la route ayant le label MPLS aboutissant vers Odéon, alors que celui du Cnam sélectionne la route dont le label MPLS aboutit à Jussieu.

Pour les flux entrants, chaque équipement exporte dans l'iBGP les routes des sites apprises du VPN avec une métrique particulière (encore l'attribut *localpref*). Sur les équipements de bordure, les annonces des routes RAP du VPN vers RENATER sont marquées avec les mêmes communautés RENATER que celles utilisées pour le service IPv4 ou IPv6, en réutilisant les mêmes règles de transformation des métriques en communautés BGP.

Du point de vue des sites, le raccordement fiabilisé à un L3VPN est réalisé en implémentant les mêmes mécanismes que ceux disponibles pour le service IPv4 ou IPv6.

Pour résumer, c'est ce mécanisme d'import/export des routes entre la VRF du L3VPN et l'iBGP sur chaque point de présence qui permet d'implémenter une ingénierie de trafic identique à celle mise en place pour le service de routage IPv4 ou IPv6.

3.3 Architecture cible

L'architecture cible représentée à la figure 9 consiste à remplacer le *peering* établi directement par chaque site avec le NR de Paris-1 par d'une part un *peering* sur chaque équipement RAP de bordure avec RENATER, et d'autre part un *peering* entre le site et son point de présence de raccordement. Les équipements de RAP apprendront les routes de SIRES annoncées d'un côté par RENATER et de l'autre par les sites et les redistribueront dans l'iBGP de RAP. L'intérêt de cette configuration est double :

- les sites bénéficieront de la redondance des accès de RAP à RENATER de manière transparente, car en cas de panne de l'accès principal au L3VPN RAP, la connexion basculera vers le second accès ;
- les sites n'auront plus qu'un seul interlocuteur pour le service de L3VPN, RAP, au lieu de deux précédemment, RAP et RENATER.

[21] *Provider Edge*

[22] L'attribut *route-distinguisher* est unique pour chaque routeur PE et chaque VPN ; concaténé à la route IPv4 ou IPv6 du VPN, il va permettre d'individualiser chaque route au sein du réseau de l'opérateur. Ainsi, il sera possible de transporter des routes identiques provenant de VPN différents (typiquement, des routes privées de la RFC 1918).

[23] *Virtual Routing and Forwarding*

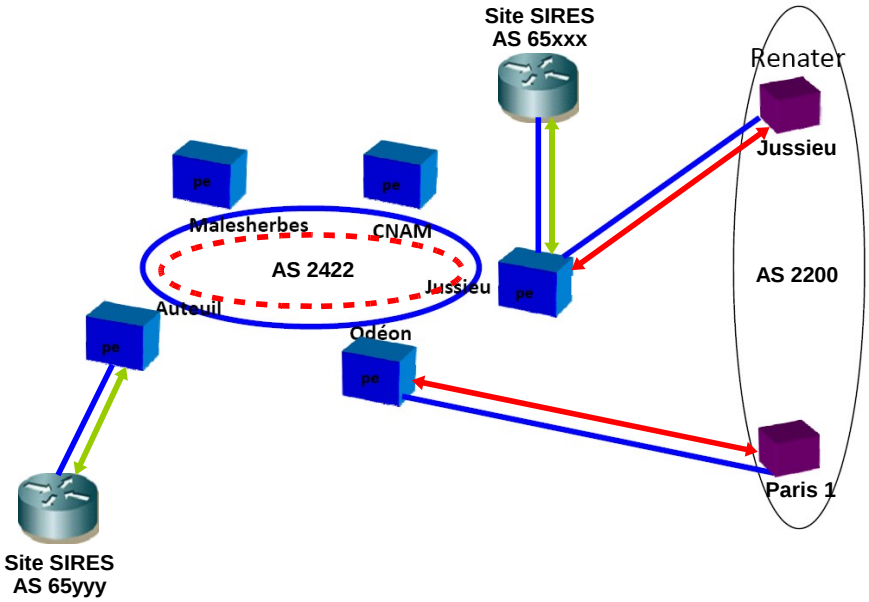

Figure 9 – *Architecture cible*

La fiabilité des raccordements au L3VPN peut encore être améliorée par un double raccordement des sites à RAP. Dans ce cas le site aura deux *peering* BGP dans deux VLAN distincts avec deux équipements de RAP, comme représenté sur la figure 10.

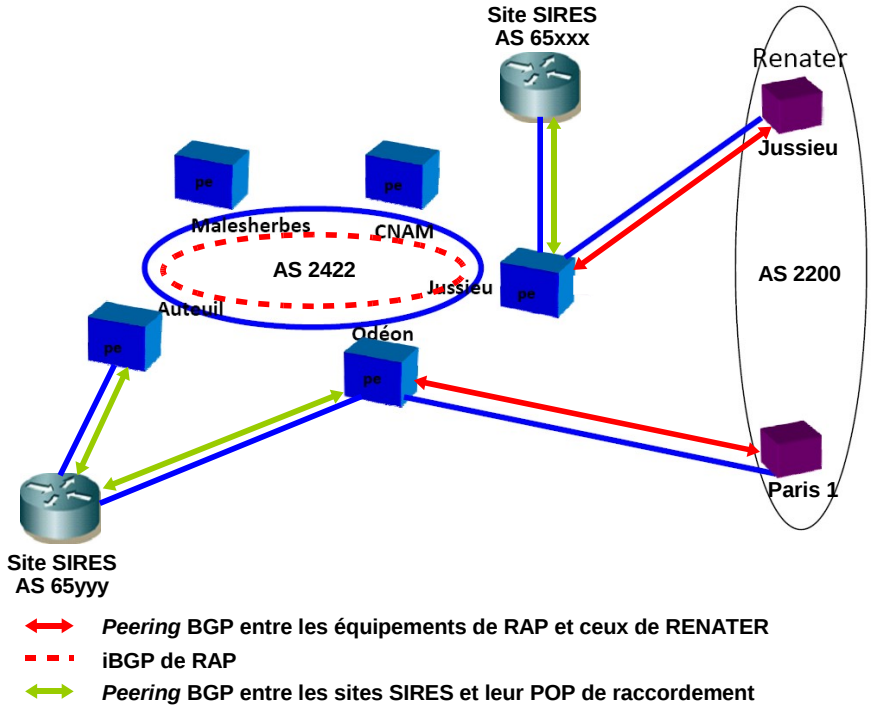

Figure 10 – *Site en raccordement fiabilisé sur RAP*

Le site annoncera chacun de ses préfixes avec la communauté 2422:100 si le chemin par le point de présence de Jussieu et le NR Jussieu est le chemin privilégié pour ce préfixe, 2422:200 si le chemin par Odéon et le NR Paris-1 est le chemin privilégié. RAP routera le trafic entrant à destination de ces préfixes par le chemin privilégié pour chacun d'entre eux. Ce mécanisme peut

fonctionner en mode actif/passif (annonce de tous les préfixes avec la même communauté) ou en partage de charge (annonce d'une partie des préfixes avec la communauté 2422:100 et de l'autre partie avec la communauté 2422:200).

Le site recevra les routes de SIRES par les deux accès. Le trafic sortant devra être routé par le chemin privilégié pour chacun des préfixes source, de façon à ce que le routage soit toujours symétrique. En mode actif/passif cela peut être fait en affectant une *local-preference* en fonction de l'interface par laquelle on reçoit les routes. En mode partage de charge il faudra router en fonction de l'adresse source.

4 Conclusion

L'architecture de routage sur RAP a évolué au fur et à mesure des besoins des sites. Elle permet de mettre en place à la demande des établissements des services avancés. La prolongation sur RAP des services L3VPN opérés par RENATER facilite la prise en compte des accès multiples disponibles pour les sites. Pour les sites parisiens du réseau SIRES, RAP devient alors l'interlocuteur unique pour l'accès au L3VPN et la gestion de celui-ci.

Bibliographie

[1] Centre Opérationnel du Réseau Académique Parisien, Mise en place de raccordements fiabilisé pour les sites sur RAP, document de spécifications disponible sur http://www.rap.prd.fr/services/ipv4.php

[2] Centre Opérationnel du Réseau Académique Parisien, Architecture pour la prise en compte des raccordements multiples, document de spécifications disponible sur http://www.rap.prd.fr/services/ipv4.php

[3] Laurent Gydé, La refonte du backbone de RAP, JRES 2009

[4] RFC 4364 BGP/MPLS IP Virtual Private Networks (VPNs)