Multihome BGP sur REVE
Retour d'experience

Eric Doutreleau
GET– INT, 9 rue charles Fourrier Évry
eric.doutreleau@int-edu.eu

Jehan Procaccia
GET– INT, 9 rue charles Fourrier Évry
jehan.procaccia@int-edu.eu

Guy Orrado
Université d'Évry, Blv F.Mitterand Évry
guy.orrado@univ-evry.fr

Claude Scarpelli
Genoscope, 2 rue Gaston Cremieux Évry
claude.scarpelli@genoscope.cns.fr

Résumé

Divers événements et besoins critiques ont introduit la nécessité pour le Réseau Métropolitain d'Évry Val d'Essonne (REVE) de disposer d'un Fournisseur d'Accès Internet (FAI) en plus de l'accès de commodité offert par RENATER. Bien que prévu dès la conception du réseau, l'introduction d'un second opérateur n'a pas été une sinécure. Au delà de l'étude de marché et de la prise en compte des contraintes imposées par les clients de REVE et les opérateurs Internet concernés, il a fallu mettre en œuvre toute une politique de routage basée sur LE protocole de routage de l'Internet : BGP. Après une phase de maquettage, suivie d'une rapide mise en production liée à un événement prévu (déménagement du Nœud RENATER), bien que parfaitement opérationnel lors de réelles défaillances d'un des deux opérateurs, ce multi-accès Internet basé sur BGP (multi-home BGP) a également généré un certain nombre de comportements imprévus. Nous détaillons d'une manière factuelle et relativement chronologique comment nous avons vécu et compris cette nouvelle architecture.

Mots clefs

BGP, REVE, RENATER, FAI, routage, multihome, Réseau Métropolitain (MAN), Réseaux de collecte.

1 Introduction

Le Réseau métropolitain d'Évry Val d'Essone (REVE) héberge un certain nombre d'utilisateurs qui ont un usage critique des ressources réseau. Il a été dès sa conception architecturé autour d'une topologie entièrement redondante. En revanche, il ne disposait initialement que d'un seul point de sortie sur l'Internet, constituant ainsi son « single point of failure ». Pour parer à cette faiblesse, mais aussi pour répondre à d'autres contraintes, nous avons doté le réseau REVE d'une nouvelle connexion physique sur un Fournisseur d'Accès Internet (FAI) en plus de l'accès à RENATER.

Après une présentation du réseau REVE, nous aborderons quelques notions du protocole de routage (BGP), employé pour la mise en place de cette double connectivité Internet. Le cœur de cet article sera ensuite axé sur le retour d'expérience que nous avons eu autour de cette mise en œuvre. Enfin, nous terminerons sur les perspectives d'évolution de cette nouvelle architecture.

2 REVE, Réseau d'Évry Val d'Essonne

2.1 Présentation du projet

Pour favoriser le développement des communications « haut débit » de leurs établissements implantés à Évry, un groupe d'une dizaine de partenaires publics et privés se sont réunis afin d'établir un réseau de télécommunication « privatif » sur le territoire d'Évry.

Les acteurs directement concernés par le projet sont : l'Université d'Évry Val d'Essonne (UEVE) ; le GIP GENOPOLE ; le Consortium National de Recherche Génomique ; l'Institut National des Télécommunications (INT), membre du Groupe des Ecoles de Télécommunications ; le Centre des Matériaux de l'Ecole Nationale des Mines de Paris (ENSMP) ; le laboratoire GENETHON ; l'Ecole Nationale Supérieure pour l'Industrie et l'Entreprise (ENSIIE). De plus, sont associés à ce projet des acteurs nationaux du monde de la recherche (comme le CNRS, l'INRA via l'unité URGV ou l'INSERM). Ces laboratoires sont intégrés dans les structures partenaires du projet du GIE REVE. Un dossier a été instruit par l'Autorité de Régulation des Télécommunications, et a abouti à l'obtention d'une licence au titre de l'article L33.2 de la loi de règlementation des télécommunications.

Le projet comporte plusieurs volets : un volet «infrastructure fibre optique » qui a été livré début 2001 dans le cadre d'une prestation dont la maîtrise d'ouvrage a été assurée par le SAN (Syndicat d'Agglomération Nouvelle d'Évry) et le suivi pour le GIE REVE par un partenaire (AURIF : Association des Utilisateurs des Réseaux d'Ile de France) mandaté par la Région pour nous accompagner pendant toutes les phases du projet. Un volet « fourniture des équipements actifs et exploitation » qui a fait l'objet d'un marché et un volet pour la fourniture d'un accès haut débit longue distance vers « Internet ». Dès le démarrage, les membres ont utilisé le réseau RENATER

via le point d'accès (NR : Nœud RENATER) implanté à Évry dans les locaux techniques du GIS INFOBIOGEN. (NR ensuite déplacé à l'université d'Évry).

Le réseau métropolitain d'Évry Val d'Essonne est opérationnel depuis le mois de juin 2001.

2.2 Fonctionnement opérationnel du projet

Le Groupement d'intérêt Economique (GIE) REVE, créé le 22 juillet 1999, a en charge la maîtrise d'ouvrage et le bon fonctionnement du Réseau Indépendant à Usage Partagé. Cette structure est présidée par le Président de l'Université d'Évry Val d'Essonne.

Un Comité Technique, chargé du choix du prestataire, du suivi de l'exploitation (relations avec l'exploitant) et des évolutions, est représenté par un membre du CNRG, un membre de l'INT et un membre de l'UEVE. La responsabilité administrative du projet est confiée à un membre de l'UEVE en charge de la mise en place et de la gestion du marché passé avec le prestataire et accompagnée par un organisme extérieur.

L'exploitation du réseau a été confiée à la société C-S à l'issue d'une procédure d'appel d'offre.

2.3 Financement du projet

2.3.1 Financement des infrastructures

Le coût de réalisation de l'infrastructure (réseau de fibres optiques non éclairées et génie civil), incluant les études amonts et l'assistance à maîtrise d'ouvrage, s'est élevé à 850 K€ TTC. Son financement a été assuré par :

- Une subvention du Conseil Régional d'Ile-de-France à destination de l'Université d'Évry, représentant 50% du montant (430 K €),

- Le SAN, qui a ensuite remis gratuitement l'ouvrage à disposition du GIE REVE, via une convention entre le SAN et l'UEVE (destinataire de la subvention du Conseil Régional).

2.3.2 Financement du réseau

Le financement du GIE est assuré par des cotisations de ses membres, appelées d'après le budget annuel prévisionnel établit par le Conseil d'Administration. Une attention particulière a été apportée au financement des améliorations du réseau (la réserve). Une faible part de cette réserve (définie dans le règlement intérieur) sert à financer des extensions géographiques.

3 Topologie du réseau REVE

3.1 Principes

L'infrastructure est constituée de trois tronçons de 72 fibres optiques monomodes interconnectées (sur panneau de brassage à trois fois 72 connecteurs) en deux points du réseau (sites en « Y » de l'Université et de Généthon). Les 11 autres sites sont installés en «dérivation» sur l'un des tronçons, et offrent la disponibilité de 12 paires de fibres sur connecteurs (les autres étant soudées). Deux paires sont exclusivement réservées au site (utilisées pour un service IP redondant), et dix paires sont communes (permettant de créer des liaisons point à point à la demande). Cette implantation permet de déployer toutes les topologies logiques (notamment des étoiles centrées sur de multiples localisations).

3.2 Plan niveau 1

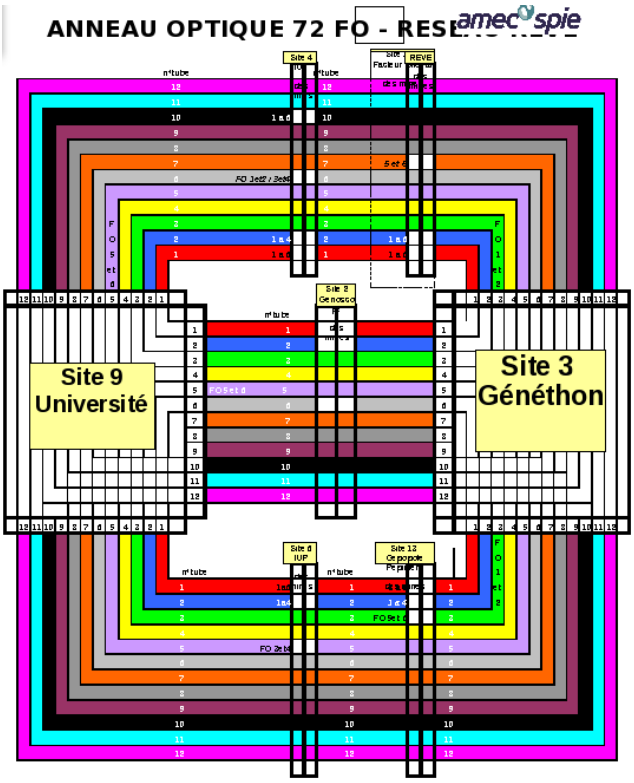

Figure 1: Plan physique simplifié du réseau du GIE REVE, sur l'agglomération d'Evry

3.3 Fonctionnement niveau 2 et 3

Le service IPv4 unicast , caractérisé par un haut débit et un faible temps de latence, est disponible sur les sites utilisateurs à travers deux paires de fibres optiques (redondance). Un commutateur Ethernet est fourni par le GIE REVE, et constitue la limite de responsabilité de l'opérateur vis à vis des utilisateurs. Les utilisateurs sont invités à connecter un routeur sur ces commutateurs. Leur trafic est alors amené via la double étoile sur le cœur du réseau IP à travers un VLAN dédié (un par routeur utilisateur).

La redondance est assurée par le protocole **Hot Standby Router Protocol** (HSRP) sur les 2 commutateurs (InfobioVI et infobioVP, cf schéma suivant) de type CISCO cat6506 au cœur du réseau REVE. Ainsi, chaque site client communiquera avec ce cœur en utilisant une adresse IP virtuelle (HSRP) de sortie. Par exemple, le site de l'INT définit une route par défaut sur l'adresse virtuelle 194.57.241.203, alors que les adresses physiques des deux commutateurs se terminent respectivement en .201 et .202 .

Pour augmenter encore la résilience et la sureté de fonctionnement du réseau IP du GIE REVE, les commutateurs et les routeurs sont installés dans deux bâtiments distincts de l'Université d'Évry.

L'accès vers les opérateurs se fait à travers deux routeurs cisco 7206, qui réalisent aussi les opérations de «traffic engineering» (essentiellement shaping) qu'impose parfois le contrat entre certains utilisateurs et certains FAI.

Chaque client de REVE annonce ses réseaux IP au travers de sessions BGP avec le cœur de réseau REVE. Les clients disposent de numéros d'AS privés sur REVE. C'est l'ensemble du réseau REVE qui dispose d'un numéro d'AS publique, (2072), afin de pouvoir s'annoncer vers plusieurs opérateurs. Exemple de l'INT :

```
router bgp 65026
network 157.159.0.0
neighbor 194.57.241.203 remote-as 2072
```

Enfin, les services IPv6 et multicast IPv4 sont, en l'attente d'une mise a jour de l'IGP de REVE (passage de OSPF à IS-IS), non redondés pour l'instant.

3.4 Schema niveaux 2 et 3 REVE

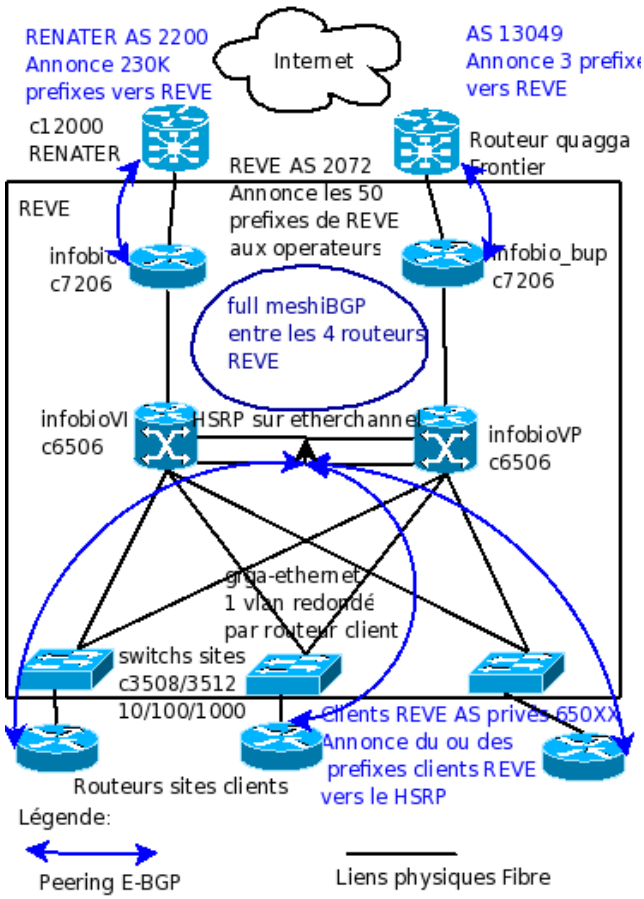

Figure 2 - niveaux 2 et 3 sur REVE

3.5 FAI et politique de routage

3.5.1 Choix du FAI initial

La totalité des premiers utilisateurs du réseau du GIE REVE était éligible à RENATER qui a donc tout naturellement été connecté prioritairement au réseau, via le point de présence installé en 2000 dans les locaux de feu le GIS INFOBIOGEN.

Cependant, la définition initiale des besoins avait fait apparaître la nécessité d'un second FAI, afin de desservir les utilisateurs potentiels du réseau non éligibles à RENATER (typiquement, l'activité commerciale des sociétés de biotechnologie installées ou à venir sur le périmètre du campus de la Génopole® d'Évry).

Le trop faible nombre de cette catégorie d'utilisateurs n'a jamais permis de dégager les conditions économiques favorables à l'installation d'un second FAI ; en particulier, le coût du simple raccordement du réseau du GIE REVE à un point de présence d'opérateur était jugé prohibitif.

Les conditions vont être radicalement modifiées entre 2005 et 2006. L'annonce de l'arrêt des services du GIS INFOBIOGEN, qui hébergeait le point de présence de RENATER dans un bâtiment loué (et pour lequel ni l'université d'Evry ni le GIP Génopole® ne prévoyaient, sur le long terme, de conserver un bail), nécessitait de trouver une solution de continuité de service pendant les opérations de déménagement. Parallèlement, plusieurs membres du GIE REVE ont exprimé le souhait de redonder l'accès Internet.

L'existence d'un point de présence de la société Frontier Online aux abords immédiats du réseau a permis de résoudre la contrainte économique du coût du raccordement du réseau du GIE REVE à un point de présence éloigné.

Grâce à ce second opérateur, le GIE REVE se trouvait en mesure d'accueillir des membres utilisateurs non éligibles à RENATER, quasiment au coût marginal de leur trafic. Mais surtout, il devenait envisageable économiquement de faire réaliser par ce second opérateur le transit de secours du trafic des membres du GIE REVE éligibles à RENATER.

Nous décrivons ici les aspects techniques de cette opération.

3.5.2 Choix de Frontier Online

Après une consultation non formalisée auprès d'opérateurs Internet (Colt, Cogent, Completel, NeufCegetel), l'opérateur local Frontier Online nous a proposé une offre favorable. Au delà de leur présence aux abords du cheminement physique des fibres de REVE, ce qui a minimisé le coût de génie civil, cet opérateur dispose d'une double connectivité (2 liens distincts entre Evry et Paris) sur Téléhouse-2, un point de présence majeur pour une connectivité Internet nationale et internationale. De plus ils échangent du trafic (peering) avec RENATER au SFINX. Enfin leur rapide mise à disposition et des tarifs concurrentiels nous ont conforté dans ce choix.

4 BGP, définitions, principes Introduction

Nous faisons ici quelques rappels « théoriques » sur BGP qui seront loin d'être exhaustifs. Pour de plus amples détails, le lecteur pourra se reporter à la bibliographie en fin d'article. Nous décrivons ici les seules notions spécifiques à notre cas d'étude.

4.1 Une définition simple

Border Gateway Protocol (BGP) est un protocole de routage qui permet d'échanger des préfixes entre différents domaines réseaux sur Internet. Dans chaque domaine administratif (ou Autonomous System : AS) un ou plusieurs équipements (routeurs) sont chargés de gérer le routage BGP. Le routage interne d'un Autonomous System est géré par un autre protocole de routage. L'avantage de BGP est qu'il peut gérer un très grand nombre de routes et qu'il permet une grande flexibilité dans la configuration de politiques de routage. Il est aujourd'hui le seul protocole permettant l'échange de toutes les routes de l'Internet (230 000 préfixes IPv4 environ)

Le routeur est en communication avec ses voisins au moyen d'une session BGP pour échanger des routes et annoncer sur internet l'existence de l'Autonomous System. A partir des routes qu'il apprend de ses voisins, il va savoir comment envoyer les données qui sont destinées à une adresse particulière.

4.2 AS-PATH

Chaque route possède un attribut nommé ***AS-PATH*** qui contient la liste des numéros d'AS au travers desquels le préfixe a été annoncé. On connait ainsi le chemin de retour de l'annonce. Mais le même préfixe peut avoir plusieurs annonces et donc il faut pouvoir sélectionner la meilleure route. On peux artificiellement « allonger » l'AS-PATH (***AS-Prepend***) afin de défavoriser un chemin.

4.3 Local-Pref

La ***l****ocal-pref* permet à un routeur de spécifier lui même quelle est la préférence qu'il accorde aux différents chemins dont-il dispose (qu'il a appris) pour joindre un même préfixe. Le choix est local et non dicté par l'annonceur du préfixe, ce qui « politiquement » a une incidence capitale. Plus la valeur attribuée à l'attribut de **local-pref** est importante, plus la route est préférée.

4.4 Algorithme de sélection de routes

Voici l'algorithme simplifié utilisé pour départager 2 routes vers un même préfixe :

1°) La route qui dispose du « Local-pref » le plus important prime. En effet chaque AS peut décider de lui même qu'il est plus intéressant politiquement d'utiliser une route plutôt qu'une autre.

2°) La route qui dispose du plus court AS-Path prime.

3°) D'autres règles interviennent alors mais elles ne concernent pas notre cas d'étude.

4.5 iBGP

Pour des raisons pratiques (redondance, multi-attachement), il est souvent nécessaire d'échanger les tables de préfixes BGP entre différents routeurs d'un même AS : cela est fait via le protocole iBGP. Il est alors nécessaire de créer une session iBGP entre chaque paire de routeur (full-mesh). On retiendra que les annonces reçues par iBGP ne sont pas retransmises aux autres peers.

4.6 Dampening

Il est assez pénalisant pour un routeur d'avoir à ajouter puis de retirer périodiquement des routes dans le cas où le routeur qui les annonce subit des dysfonctionnements intempestifs. Pour éviter ces « bagottements » (route flap), certains routeurs sont configurés pour faire du BGP ***dampening,*** c'est-à-dire affecter une pénalité aux routes en fonction de la stabilité de l'annonce qu'ils reçoivent. Ces routeurs feront alors disparaître ou réapparaître une route suivant que la pénalité atteint ou redescend en dessous d'un seuil défini Il s'agit d'une propriété optionnelle du protocole BGP, qu'un opérateur pourra mettre en place s'il désire « stabiliser » ses annonces.

5 Multihome BGP sur REVE

5.1 Rappel du besoin

Au delà de la nécessité de backup en cas de pannes du NR d'Evry, un second FAI permet également aux clients et trafics de REVE non éligibles RENATER de disposer d'un accès Internet mutualisé. Enfin un partage de charge est aussi envisageable.

5.2 Contraintes

Des contraintes ont été imposées par les utilisateurs REVE et l'opérateur RENATER :

[1] non renumérotation des réseaux des membres actuels

[2] annonce de préfixes RENATER via Frontier

[3] priorité RENATER en sortie et entrée pour les sites éligibles

[4] éviter un routage asymétrique

[5] routage politique dans REVE

5.3 Solution retenue

Après avoir évoqué différentes techniques afin de répondre aux contraintes ci-dessus (communauté BGP, local-pref, AS-prepend), le choix s'est porté sur de l'AS-prepend car la principale contrainte était de favoriser le trafic aller et retour via RENATER pour les sites éligibles. Le principe a été de répéter artificiellement trois fois de suite dans l'AS-PATH des routes annoncées à Frontier l'AS de REVE (2072). Ainsi, les routeurs de l'Internet devraient trouver le chemin de retour sur REVE via Frontier plus « long » que par RENATER. Exemple initial depuis le Looking Glass (LG) de Gitoyen sur un préfixe du Genoscope où l'on voit bien l'AS prepend apparaître derrière l'AS de Frontier (13049) :

```
show ip bgp 195.83.222.0
BGP routing table entry for 195.83.222.0/24
Paths: (2 available, best #2, table Default-IP-
Routing-Table)
  Advertised to non peer-group peers:
  80.67.168.22 80.67.168.4 80.67.168.18
80.67.168.19
  13049 2072 2072 2072 2072
    194.68.129.191 from 194.68.129.191
(64.210.69.254)
      Origin incomplete, localpref 1000, valid,
external
      Community: 20766:21000
      Last update: Thu Jul 20 19:03:26 2006
  2200 2072
    194.68.129.102 from 194.68.129.102
(193.51.178.11)
      Origin incomplete, localpref 1000, valid,
external, best
      Community: 20766:21000
      Last update: Thu Jul 20 17:23:15 2006
Configuration BGP sur le routeur de bordure
REVE<->Frontier (infobio_bup)
route-map V4-REVE_FrontierOnline-OUT permit 10
 match ip address prefix-list V4-REVE
 set as-path prepend 2072 2072
```

5.4 Forcer le retour par RENATER

Évidement, la technique d'**AS-Prepend** n'empêche pas certains FAI de choisir un chemin autre que celui basé sur la longueur de l'AS-PATH. Ils peuvent (suivant leur accord de peering/politique commerciale etc ..) positionner des **local-pref** qui préféreront Frontier à RENATER malgré un AS-PATH défavorable. Pour contourner cette incertitude vis à vis du choix des autres FAI et éviter un routage asymétrique, nous avons convenu de faire en sorte que Frontier annonce les préfixes de REVE uniquement lorsque le lien entre REVE et RENATER devient inopérant et/ou que les annonces de préfixes REVE via RENATER ne sont plus visibles. Ainsi ces autres FAI ne verront à tout instant qu'un seul chemin de retour sur REVE.

Pour ce faire, Frontier a profité d'un comportement distinct vis à vis de la réannonce de routes, entre des sessions iBGP et eBGP au sein de leur backbone. Si Frontier ne reçoit plus les préfixes de REVE via le peering qu'ils ont au SFINX avec RENATER, ils se mettent alors à annoncer les préfixes de REVE au reste de l'Internet, et ainsi nous assurent une continuité de service quand RENATER est inopérant vis à vis de REVE. Autrement, quand ils reçoivent les routes de REVE via le SFINX, ils n'annoncent rien concernant REVE et alors nous ne sommes visibles des autres ISP que via RENATER. On évite donc un routage asymétrique malgré l'AS-prepend.

Le détail est expliqué sur la figure ci-contre : tout ce joue sur « core1 ». Quand le peering entre Frontier et RENATER sur le SFINX est opérationnel, « edge3 » affecte une Local-Pref (LP) de 800 aux routes de REVE. « core1 » reçoit les routes de REVE via la session iBGP de « edge3 » (LP800) et la session eBGP (LP700) depuis le peering avec REVE. Dans ce cas il n'annonce pas à « egde2 » (donc l'Internet) les préfixes de REVE car la meilleure annonce (LP800) est venue d'une session iBGP (iBGP est intransitif : on ne diffuse pas en iBGP de routes apprises par iBGP). En revanche si le peering entre Frontier et RENATER sur le SFINX tombe, alors « core1 » ne dispose plus que de l'annonce des préfixes de REVE via la session eBGP (LP700). Comme il s'agit de préfixes appris ici dans le cadre d'une session eBGP, donc « réannonçables », « core1 » diffuse l'annonce à « edge2 » et ce dernier sur l'Internet . Les préfixes de REVE sont donc annoncés par Frontier.

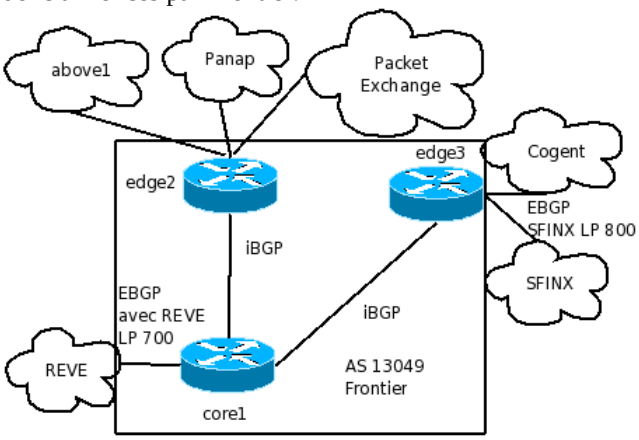

Figure 3: Routage sur Frontier

6 Etapes de mis en œuvre

6.1 Full routing sur REVE

Le routeur de bordure (« edge ») de REVE reve-infobio c7206 reçoit le 20 Juillet 2006 l'annonce de toutes les routes connues de l'Internet depuis RENATER (Full routing). Celles-ci sont propagées en iBGP sur les autres routeurs du cœur de réseau REVE (cf topologie/schéma). Les clients REVE positionnent LA route par default (**0.0.0.0/0**) qui pointe sur le cœur de réseau REVE. Pour des raisons de redondance cette adresse est virtuelle (HSRP).

6.2 Tests préalables

Avant de passer en production, à savoir basculer réellement sur le lien de backup (Frontier) à l'occasion du déménagement du Nœud RENATER d'Evry prévu le 29 Juillet 2006, nous avons simulé une défaillance de RENATER en coupant le peering entre le NR d'Evry et le routeur de bordure de REVE. Après quelques minutes

(entre 5 et 10 suivant les préfixes à joindre), le temps que les annonces se propagent sur l'Internet, Frontier prend bien le relais. Ci-dessous un exemple depuis le LG de Gitoyen pour le préfixe de l'INT (à noter l'AS-prepend) :

```
BGP routing table entry for 157.159.0.0/16
Paths: (2 available, best #1, table Default-IP-
Routing-Table) Not advertised to any peer
13049 2072 2072 2072 2072
80.67.168.4 from 80.67.168.4 (80.67.168.4)
Origin IGP, localpref 1300, valid, internal,
best Community: 20766:23000
Last update: Thu Jul 20 19:19:20 2006

Traceroute
[jehan@calaz.int-evry.fr ~]$ traceroute
rigolo.nic.fr
...
3  int-50-h.reve.fr (194.57.241.203)  0.392 ms
0.378 ms 0.376 ms
 ...
6  core2.th2.frontier.fr (213.161.200.17)  1.853
ms 1.984 ms   1.781 ms
7  edge3.th2.frontier.fr (213.161.200.23)  1.733
ms 1.800 ms   1.835 ms
8  Renater.sfinx.tm.fr (194.68.129.102)  2.449
ms   2.097 ms   2.060 ms
 ...
12  rigolo.nic.fr (192.134.4.20)  3.333 ms
3.127 ms   3.255 ms
```

6.3 Événements imprévus

Lors de cette première bascule, certains événements imprévus nous ont interpelés.

6.3.1 Traitements particuliers sur les "gros" préfixes

Le FAI **above.net,** dont est client Frontier, devait filtrer les annonces de « gros » préfixes. Cela a été le cas pour la classe B[1] de l'INT (157.159.0.0/16) qui ne pouvait joindre une partie de l'Internet (derrière Above) jusqu'à ce que les administrateurs de Frontiers demandent à Above d'ouvrir leur filtrage sur la classe B de l'INT. Par ailleurs le temps de convergence, notamment pour la classe B de l'INT a été relativement long, environ une demi-heure pour joindre à nouveau certains domaines comme abc.com, sendmail.org, google.fr. Ce délai était probablement dû à un ***dampening*** (cf 4.6) du préfixe de l'INT lié aux différents allers-retours (**flaps**) RENATER/Frontier lors de nos tests. Les autres clients REVE disposant de classes C agrégées par RENATER sont moins, voire pas du tout sujets au ***dampening.***

Exemple de table mentionnant du ***dampening*** sur la classe B de l'INT lors de nos tests initiaux (21/07/06)

```
Nri-b Looking Glass Results
Command: show ip bgp ipv4 unicast 157.159.0.0
....
2200:4000
Originator: 193.51.178.64, Cluster list:
195.220.98.17 13049 2072 2072 2072, (suppressed
due to dampening)
194.68.129.191 from 194.68.129.191
(64.210.69.254) Origin IGP, localpref 410,
valid, external Community: 2200:2000
Dampinfo: penalty 1442, flapped 3 times in
00:19:00, reuse in 00:03:14
...
```

[1]Pour des raisons pratiques nous utilisons l'ancienne terminologie (non CIDR)

6.3.2 Sous dimensionnement d'un routeur

Le routeur edge de REVE connecté à Frontier (infobio_bup, CISCO 7206 avec une NPE400 et 256 de RAM) n'a pas supporté l'annonce de toutes les routes (190K) de l'Internet. Après seulement 12K routes, la CPU était à 100%, il s'en est suivi une perte des peering BGP et un début d'annonces de routes depuis Frontier qui semblait également perturber RENATER en interne. On abandonne alors le full routing entre Frontier et REVE, tout rentre ensuite dans l'ordre. Le routeur de bordure REVE-Frontier ne reçoit alors plus que trois routes de Frontier (les 2 préfixes de Frontier et une route par défaut) et envoie à Frontier la cinquantaine de préfixes de REVE, ce qui reste très léger.

6.3.3 Abandon du "Shortest Path" REVE - Frontier

Initialement, Frontier a positionné un ***local-pref*** sur les annonces de REVE qui favorisait le trafic des clients de Frontier via le lien de backup (REVE-Frontier). Afin de respecter les termes de notre accord entre RENATER et Frontier, ce dernier a mis un local-pref défavorable (300) sur les préfixes de REVE vers le lien de backup et un plus favorable (800) vers le SFINX-RENATER. Ainsi cela force le trafic entre REVE et Frontier via RENATER et cantonne ce lien à son simple rôle de backup.

6.4 Bascule Réelle

Une bascule réelle a lieu le vendredi 28 Juillet 2006 après-midi afin de couvrir le temps de déménagement du Nœud RENATER prévu pour le weekend. Après 6 à 7 minutes de « black-out » (REVE est coupé de l'Internet), nous annulons cette première tentative de backup. La bascule aboutit enfin avec un peu plus de patience, lors de la deuxième tentative. L'opérateur présume que le premier échec de bascule est lié à infobio_bup qui supporte difficilement la charge au moment de la bascule. A cette deuxième tentative, après 8 minutes d'attente de convergence BGP, les clients REVE disposent enfin d'une gateway opérationnelle vers l'Internet via infobio_bup puis Frontier :

```
[jehan@calaz.int-evry.fr ~]$ traceroute
www.imag.fr
...
3  int-50-h.reve.fr (194.57.241.203)  0.458 ms
0.379 ms 0.623 ms
4  195.83.166.4 (195.83.166.4)  112.261 ms
111.837 ms 109.898 ms
5  core1-0-2.evry1.routers.frontier.fr
(213.161.192.65) 0.576 ms   0.583 ms 0.499 ms
6  edge3.th2.routers.frontier.fr
(213.161.200.23)  1.427 ms 1.449 ms 1.333 ms
7  Renater.sfinx.tm.fr (194.68.129.102)  2.013
ms   2.055 ms   1.842 ms
 ...
13  rillette2.imag.fr (129.88.34.211)(H!)
13.581 ms (H!) 13.556 ms (H!) 13.477 ms
```

Vision de l'AS de REVE sur le LG Gitoyen à ce moment :

```
Executing command = show ip bgp regexp 2072 BGP
table version is 0, local router ID is
80.67.168.16 Status codes: s suppressed, d
damped, h history, * valid, > best, i – internal
Origin codes: i – IGP, e – EGP, ? - incomplete
...
•       157.159.0.0      194.68.129.201
21   1000      0 6461 13049 2072
•       2072 2072 i

Total number of prefixes 51 giga-bgpd
```

Ci-dessous la charge CPU constatée sur infobio_bup 10 minutes après la bascule. On voit que le routeur a eu des difficultés à supporter la bascule :

```
infobio_bup#sh processes cpu
CPU utilization for five seconds: 15%/14%; one
minute: 20%; five minutes: 42%
```

Autre symptôme curieux, depuis l'Internet et/ou Frontier vers l'INT, nous avons constaté lors d'une bascule des changements irréguliers de routage interne entre les deux 6506 du cœur de réseau REVE (infobioVI et infobioVP) avec des pertes de paquets allant jusqu'à ~50% , cf mtr :

```
My traceroute  [v0.71] calaz (0.0.0.0) Tue Nov  7
15:54:02
2006

Packets               Pings
Host                                        Loss%  Last
Avg  Best  Wrst StDev
....
3.      int-50-h.reve.fr                     0.0%
1.7   1.4   0.4 298.0 25.6
4.      195.83.166.4                        22.4%
1.6  13.9   0.6 1549.  54.6
5.      c7204-ext.reve.fr                   44.7%
1.6   5.8   0.5 299.9  29.0
6.      evry-a3-0-3.cssi.renater.fr          0.0%
2.5   1.2   0.5 121.2   4.8
7.      orsay-g5-1.cssi.renater.fr          44.7%
117.0  13.5 1.9 330.5  36.1
....
10.     251.renater.fr                      44.7%
8.1   3.5   2.5 289.9 7.3
    11. www.renater.fr                       0.0%
        5.7   2.9 2.0 123.8 5.2
```

Ce sont les prémisses d'une longue série de pertes de connectivités intempestives liées à des boucles de routage internes à REVE. Pour y remédier, sans pour autant en avoir vraiment déterminé la cause, l'opérateur force un routage interne à REVE statique pour joindre le site de l'INT via infobioVI (un des 6506, cf shéma), mais on perd alors tout l'intérêt de la haute disponibilité du double cœur de réseau REVE. Exemple avec l'INT (vlan 50) pour lequel on force une priorité HSRP sur infobioVI :

```
infobiovi#sh standby vlan 50
Vlan50 – Group 50
Local state is Active, priority 105, may preempt
```

7 Etape de production

Après l'opération de déménagement du NR d'Evry, la configuration de routage sur REVE reste ainsi en continu sur ce mode de ***multihome*** BGP avec RENATER en principal et Frontier en backup. Hélas, les pertes de connectivité constatées lors de la première bascule réelle (fin juillet 2006) vont s'intensifier au fil du temps.

7.1 Perturbations sur le Man REVE

7.1.1 Constats

Depuis la mise en place du lien de backup, des effets de bord ont plus ou moins perturbé la connectivité à l'Internet du site de l'INT[2]. Cela s'est traduit par quelques minutes de pertes durant l'automne 2006, à de sérieuses perturbations (plusieurs heures de dysfonctionnement) durant l'hiver 2006/2007. La difficulté résidait dans le fait que les pertes de connectivité n'étaient pas franches, seulement certains sites étaient injoignables, et depuis l'extérieur certaines machines de l'INT étaient visibles, d'autres pas. Les symptômes étaient vraiment déconcertants et laissaient penser plus à un problème de bug logiciel ou matériel que d'une explication logique liée à une mauvaise configuration. Cela explique la longue période de résolution du problème, marquée par nos différents tests, analyses , interventions de l'opérateur par le biais de mises à jour d'IOS, de changements de cartes, de modifications des configurations dans un environnement de productions etc...

7.1.2 Boucles de routage

Voici un exemple de boucle de routage, allers-retours entre 195.83.166.2 et 194.57.241.201. Un ***traceroute*** depuis une freebox vers le DNS primaire de l'INT :

```
[root@localhost ~]# traceroute -n 157.159.10.12
traceroute to 157.159.10.12 (157.159.10.12), 30
hops max, 40 byte packets 1  192.168.1.1
(192.168.1.1)  2.224 ms   2.037 ms 2.880 ms
....
13  193.51.181.221 (193.51.181.221)  44.039 ms
42.969 ms   43.461 ms
14  195.83.166.2 (195.83.166.2)  148.146 ms
139.951 ms 134.772 ms
15  194.57.241.201 (194.57.241.201)  43.977 ms
43.600 ms   44.255 ms
16  195.83.166.2 (195.83.166.2)  128.043 ms
127.757 ms 126.562 ms
17  194.57.241.201 (194.57.241.201)  43.275 ms
44.456 ms   43.963 ms
etc...

ping 157.159.10.12
PING 157.159.10.12 (157.159.10.12) 56(84) bytes
of data.
From 195.83.166.2 icmp_seq=1 Time to live
exceeded From 195.83.166.2 icmp_seq=2 Time to
live exceeded
```

Le fait que l'adresse IP du DNS primaire de l'INT soit affectée a largement contribué à l'aggravation de la situation. Finalement, une bonne partie des serveurs étaient joignables, mais ne pouvaient être résolus. En effet, au même moment un ***traceroute*** sur l'adresse IP d'un autre serveur de l'INT fonctionne correctement !

De plus, bien que disposant d'une supervision de l'état de connectivité de notre site par l'opérateur, son NOC (directement connecté dans RENATER) ne rentrait pas dans la boucle de routage : RENATER faisant partie des

[2]Nous n'expliquons toujours pas pourquoi seul l'INT a été perturbé, . hormis le fait que l'INT soit le seul réseau de classe B

sites non perturbés, il ne constatait donc pas le problème qui dépendait alors de la source du ***traceroute***.

Pour palier en urgence le problème, en plus de forcer le HSRP au niveau vlan sur l'un des CISCO cat6506 du cœur de réseau REVE, le NOC-REVE a mis en place un routage statique pour joindre l'INT. Cela a résolu provisoirement le problème, mais on a perdu alors tous les bénéfices de la redondance apportée par le MAN. Une application graphique (BGPlay, http://www.ris.ripe.net/bgplay/) fournie par le RIPE nous a permis de constater les incessants allers-retours de nos annonces :

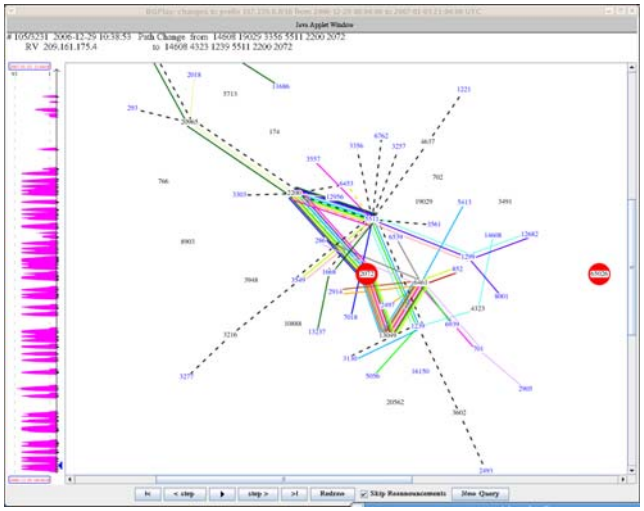

Figure 4 - *Visualisation graphique des changements de chemin avec bgplay.*

Chaque changements apparaît sur l'echelle de temps verticale à gauche en violet (il y en a anormalement beaucoup). Sur la carte on distingue que l'AS 2072 en rouge au centre, est joint suivant les AS sources, soit par Frontier, soit par RENATER.

Une première analyse du NOC a permis de constater qu'un ***load balancing*** s'était mis en place à cause de coûts OSPF (IGP de REVE) équivalents entre les deux 6506. Cela peut expliquer peut-être une partie des problèmes, mais pas tous.

7.1.3 Piste sérieuse

Il aura fallu attendre une longue période avant intervention du TAC CISCO pour donner enfin une explication assez logique aux problèmes constatés. Il semblerait en effet, qu'au fil du temps, avec l'augmentation du nombre de routes sur l'internet (de 190K en juillet à 210K en Décembre) les switch/routeur c6506 du cœur de réseau REVE n'ont plus supportés le nombre de routes à traiter, dixit le TAC CISCO:

« By default sup720 has maximum HW routes set to 192k – this is for sure not enough for the whole internet – so what happens when an entry cannot be installed in HW, traffic for this prefix coming in is sent to closest route (default GW usually) so a loop is created. »

Exemple de d'état de la TCAM[3] et du CEF[4] sur un cat6500 REVE :

```
infobiovp#sh mls cef hardware module 5 CEF TCAM
v2:
Size: 262144 entries
65536 rows/device, 4 device(s) 32 entries/mask-
block
8192 total blocks (32b wide) 1212416 s/w table
memory

infobiovp#sh mls cef summary Total routes:
194656
Ipv4 unicast routes:          194589
```

Entre temps, au delà de ces problèmes, deux réelles coupures RENATER ont eu lieu les 19 (24 minutes) et 23 (45 minutes) Janvier 2007, et le lien de backup a alors rempli parfaitement son rôle .

7.2 Disparition des boucles de routage

Mystérieusement le 30 Janvier 2007, les bascules BGP qui n'ont cessé de se produire (cf BGPlay) depuis plusieurs semaines disparaissent ! BGPlay montre bien « l'événement » en jouant une requête sur 157.159.0.0/16 entre le 30/01/2007 11H22 et 1/02/2007 19H22). Est-ce à ce moment que l'opérateur a augmenté la taille maximum de la table ***CEF*** ? (***mls cef maximum-routes ip 230***), ou bien que les pertes de connectivité du peering Frontier RENATER au SFINX se sont estompées (cf chapitre 7.3) ?

Autres événements :

Mercredi 28/02, nous surveillons une nouvelle bascule sur le lien de backup car une maintenance est programmée sur le routeur REVE-RENATER (Au final elle ne durera que 10 minutes ce qui comparé aux temps de convergences BGP constatés, aurait été peut-être moins pénalisant qu'une bascule ...) Le lien de backup joue à nouveau parfaitement son rôle. Enfin, une autre bascule (plus longue, entre autre à cause d'un problème d'accessibilité au NR) a lieu le 12 Mars suite à une saturation mémoire du NR d'Evry. De nouveau nous basculons sur Frontier.

Finalement, ce nouveau lien de backup fonctionne parfaitement bien lors de réelles pannes. En revanche, l'architecture redondante du cœur de réseau REVE, cumulée à sa faible capacité de support du full-routing, nous perturbe durant les phases de bon fonctionnement des deux opérateurs. Bien que relativement fâcheux, ces dysfonctionnements nous ont appris (empiriquement[5]) beaucoup sur le sujet, et permettent ainsi d'envisager de sérieuses perspectives d'amélioration.

7.3 Perturbations sur le peering au SFINX

Frontier est présent au SFINX, point d'échange de l'Internet opéré par le GIP RENATER.. Un peering a pu

[3]TCAM : Ternary Content Addressable Memory

[4]CEF : Cisco Express Forwarding

[5]La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. Ici, nous avons réuni théorie et pratique : rien ne fonctionne... et personne ne sait pourquoi ! (attribué à Albert Einstein)

être monté entre les deux opérateurs. Hélas, quelques pertes de connectivité de ce peering provoquent des bascules non voulues sur le lien de backup (suivant le principe expliqué en 5.4). En effet, la configuration décidée/négociée en juillet 2006 était que Frontier n'annonce les préfixes de REVE que lorsque que ce peering au SFINX tombe. Mais lorsqu'il y a perte de ce peering alors que RENATER continue par ailleurs d'annoncer nos préfixes (le NR d'Evry est UP!), les préfixes de REVE se trouvent naturellement annoncés par nos deux opérateurs. Il s'en suis alors un routage asymétrique, ce que nous voulions éviter.

Pour expliquer ces pertes de peering, les opérateurs émettent alors des doutes d'une part sur le dialogue d'une implémentation BGP **CISCO** avec une implémentation BGP libre (**quagga)**. En attendant un problème administratif coupe court à l'analyse de ce problème technique, car Frontier est déconnecté du SFINX pour manquement administratif. Le règlement de ce problème administratif devrait résoudre cette asymétrie non désirée pour autant que les dysfonctionnements techniques ne réapparaissent pas.

8 Évolutions

Bien qu'ayant montré son efficacité à l'occasion des pertes de connectivité programmées ou non de RENATER, la configuration actuelle de backup sur le second opérateur est loin d'être définitive. D'autres techniques méritent d'être étudiées, comme les communautés BGP et un principe continu et naturel de routage asymétrique. Enfin, pour réduire le facteur d'échelle lié à iBGP qui implique un réseau maillé de peering BGP (full mesh) au sein des quatre routeurs du cœur de réseau REVE, il serait souhaitable pour disposer d'un double full-routing (RENATER et Frontier), de mettre en place des ***Route Reflectors*** ou une communauté d'AS. Quoi qu'il en soit, il faudra essayer de décharger les commutateurs cat6506 du cœur de réseau REVE de cette lourdre charge de gestion de routage niveau 3 (full-routing BGP) et les cantonner à leur spécialité : le forwarding.

Annexe

Détails des événements sur :

- http://www-public.int-evry.fr/~procacci/mytwiki/bin/view/Documentations/PbRoutageReve
- http://www-public.int-evry.fr/~procacci/mytwiki/bin/view/Documentations/ReveBackupBgp

Bibliographie

[6] John W. Stewart III, BGP4 Inter-Domain Rounting in the Internet. Addison-Wesley isbn 0-201-37951-1: 1999.

[7] Orrado G., Scarpelli C. REVE, Réseau d'Evry Val d'Essonne. Dans *Actes du congrès JRES2001, pages 577-584, Lyon, décembre 2001.*

[8] Stéphane Bortzmeyer, Routage dynamique avec BGP, *http://www.generic-nic.net/formation/routage-dyn/bgp/*

[9] Jehan Procaccia, principes BGP, *http://www-public.int-evry.fr/~procacci/mytwiki/bin/view/Documentations/PrincipesBgp*