# Mise en œuvre de la « fast-convergence » dans le réseau RENATER et adaptations pour les réseaux de collecte et réseaux de campus

Franck Simon
GIP RENATER
c/o ENSAM – 151 Boulevard de l'Hôpital – 75013 PARIS
Franck.Simon@renater.fr

Florence Picard
Communications & Systèmes
22 Avenue Galilée – 92350 PLESSIS-ROBINSON
Florence.Picard@c-s.fr

**Résumé**

*L'objet de cet article est de présenter les résultats de l'étude et du déploiement qui a été mené sur le backbone national RENATER afin d'optimiser les temps de convergence des protocoles de routage dynamique (re-calcul des chemins) dans un réseau maillé.*

*Après une présentation du contexte et des raisons à l'origine d'une telle étude, cet article passe en revue l'ensemble des couches protocolaires mises en place sur le réseau RENATER et les optimisations apportées pour chacune d'entre elles.*

*Les éléments de cette étude sont ensuite repris pour introduire les optimisations qui pourraient alors être apportées au niveau de l'interconnexion avec les réseaux de collecte (réseaux métropolitains et réseaux régionaux raccordés sur les Nœuds RENATER), et au sein de ces réseaux. L'objectif recherché est alors d'avoir la meilleure convergence possible sur l'ensemble de la chaîne.*

**Mots clefs**

Fast-convergence, IS-IS, BGP, BFD, RENATER.

## 1 Introduction

Le réseau national RENATER, dans sa phase actuelle, est fortement maillé, ce qui permet de garantir un haut niveau de disponibilité : en cas d'incident ou de maintenance sur les liaisons, un re-routage automatique du trafic est effectué grâce aux nombreuses boucles constituant le réseau.

Déployer néanmoins une infrastructure maillée ne suffit plus pour garantir un excellent niveau de disponibilité : les usages et applications ont évolué au cours des années et il est nécessaire d'optimiser au maximum le temps de convergence dans le réseau, c'est-à-dire le temps de re-calcul d'un chemin en cas de défaillance d'une liaison ou d'un nœud.

Ces optimisations peuvent, au premier abord, apparaître comme des points de détails, et pourtant les améliorations sont vraiment significatives et contribuent à obtenir une architecture toujours plus robuste (contraintes de disponibilité et qualité de service croissantes).

Le but de cet article est donc de faire le point sur les principales optimisations qui ont été apportées sur le réseau RENATER, et expliquer comment cela s'est matérialisé techniquement en termes de configuration sur les équipements, et de fournir aux acteurs des réseaux de collecte des éléments pour mettre en œuvre de telles optimisations dans leur réseau.

## 2 Contexte

### 2.1 Comment la convergence est-elle mesurée dans le backbone RENATER ?

Des sondes, permettant d'effectuer des mesures actives, sont déployées dans les NR. Ces sondes fournissent, entre autres, des données précises quant à des métriques telles que le temps de transit unidirectionnel, la gigue (variation du délai), les éventuelles pertes de paquets, le nombre de sauts (nombre de routeurs traversés). Ces sondes sont équipées d'antennes GPS garantissant une synchronisation temporelle très fine. Les sondes « discutent » entre elles à raison d'environ 100 mesures par seconde.

En cas de maintenance ou d'incident générant une indisponibilité d'une des liaisons du backbone ou d'un équipement, le trafic IP est re-routé. Les données issues des sondes permettent de détecter ce re-routage, lequel a un impact direct sur :

- le délai (car nouveau chemin emprunté) même si cette variation peut être très faible (variation de une à quelques millisecondes),
- le nombre de sauts (nombre de NR traversés) associés au nouveau chemin,
- le nombre de paquets perdus (ce dernier permettant alors d'évaluer le temps de convergence de la couche IP).

## 2.2 Pourquoi mettre en œuvre une politique de « fast-convergence » ?

Les mesures collectées par les sondes ont ainsi mis en évidence des re-routages assez fréquents au sein du backbone, et qui ne sont pas forcément liés à des coupures franches ou encore des maintenances, mais aussi par exemple à des micro-coupures générées par un problème de qualité sur un lien optique.

Les sondes ont également permis de confirmer que les temps de convergence au niveau de la couche IP étaient de l'ordre de plusieurs secondes. Ce temps n'est pas forcément pénalisant pour des applications de transferts de fichiers basées sur TCP, des applications de messagerie, ou encore pour un navigateur WEB, mais peut le devenir pour des applications fortement interactives ou temps réel telles que la visioconférence ou encore la ToIP. Ces applications, dont l'usage s'est généralisé, peuvent certes être priorisées dans le réseau grâce à l'utilisation des classes de services déployées sur RENATER. Cependant, aucune priorisation, quelle qu'elle soit ne résoudra les perturbations, voire les coupures générées par un temps de convergence trop long dans le réseau.

On comprend donc que disposer d'un cœur de réseau maillé avec des équipements redondés dans les NR, garantit un haut niveau de disponibilité globale (supérieur à 99,95%), mais que cela ne suffit pas pour être « transparent » pour certaines applications.

A cet effet, le GIP RENATER, en collaboration étroite avec le NOC-RENATER, a lancé fin 2006 une étude approfondie pour optimiser les temps de convergence dans le réseau RENATER, pour un déploiement effectif en mars 2007.

# 3 Protocoles utilisés dans le réseau RENATER

Avant d'aborder les améliorations apportées dans le cadre de la mise en œuvre de la politique de « fast-convergence », il est nécessaire d'identifier les protocoles utilisés dans le réseau RENATER.

## 3.1 Le protocole de routage interne : IS-IS

Le protocole IGP (Interior Gateway Protocol) utilisé dans RENATER est IS-IS (Intermediate System to Intermediate System).

IS-IS est activé sur la totalité des liaisons de cœur de réseau, en mode topologie unique (même topologie pour IPv4 et IPv6). Avec IS-IS, une métrique est associée à chaque liaison, donc en cas de rupture de l'une d'entre elles, IS-IS recalcule le chemin réseau à emprunter en considérant la somme de ces métriques. L'algorithme utilisé pour ce calcul du meilleur chemin est SPF (Shortest Path First – algorithme Dijkstra).

Les préfixes IP diffusés dans la table IS-IS sont essentiellement les réseaux d'interconnexion associés au raccordement des liaisons physiques et logiques sur les routeurs des NR (liaisons de cœur de réseau, mais aussi liaisons des sites ou réseaux de collecte raccordés sur ces NR) : cela représente aujourd'hui environ 400 préfixes.

## 3.2 Le protocole de routage externe : BGP

Le protocole EGP (Exterior Gateway Protocol) utilisé dans RENATER est BGP (Border Gateway Protocol), dont il convient de distinguer les usages iBGP (internal BGP) et eBGP (external BGP).

### *3.2.1 Utilisation de eBGP*

eBGP est utilisé pour l'interconnexion du réseau RENATER avec les réseaux « extérieurs », c'est-à-dire entre l'AS (Autonomous System) de RENATER et tout autre AS.

Un peering eBGP est établi entre chaque NR et chaque réseau de collecte ou site qui y est directement raccordé. Dans la pratique, cela permet de recevoir de la part des éléments raccordés la liste des réseaux (préfixes IPv4 et IPv6), et de leur transmettre soit une route par défaut, soit les routes nationales RENATER, ou encore du « full-routing » (toutes les routes de l'Internet), en fonction du besoin et de la politique de routage à mettre en œuvre.

L'IGP ayant connaissance de tous les réseaux d'interconnexion de RENATER (cf. § 3.1), son rôle est d'indiquer aux routeurs RENATER comment joindre ce peering eBGP.

### *3.2.2 Utilisation de iBGP*

iBGP est utilisé pour la diffusion des préfixes au sein de l'AS RENATER, notamment pour la diffusion des préfixes IP transmis par les réseaux de collecte via les peerings eBGP établis entre ces derniers et les NR (cf. § 3.2.1). Ainsi, les préfixes appris via un peering eBGP établi entre RENATER et un réseau de collecte, sont automatiquement rediffusés dans l'iBGP, afin que chaque routeur de NR dispose du même niveau d'information dans la table de routage.

A noter que iBGP est non transitif : un préfixe appris par un routeur via une session iBGP ne sera pas ré-annoncé à un voisin iBGP. Cette règle impose un maillage logique complet (« full-mesh ») entre routeurs iBGP pour être sûr que tous disposent du même niveau d'information : chaque routeur iBGP doit alors voir l'ensemble de ses autres voisins.

Une alternative au maillage complet iBGP est d'utiliser le mécanisme de réflecteur de routes BGP. Dans le réseau RENATER, pour des raisons de redondance, deux « route-

reflectors » ont été configurés : chaque routeur de NR établit simplement une session iBGP avec chacun de ces réflecteurs. Cette solution garantit que chaque routeur iBGP dispose du même niveau d'information, tout en réduisant le nombre de sessions iBGP à établir (et donc à superviser) pour cela.

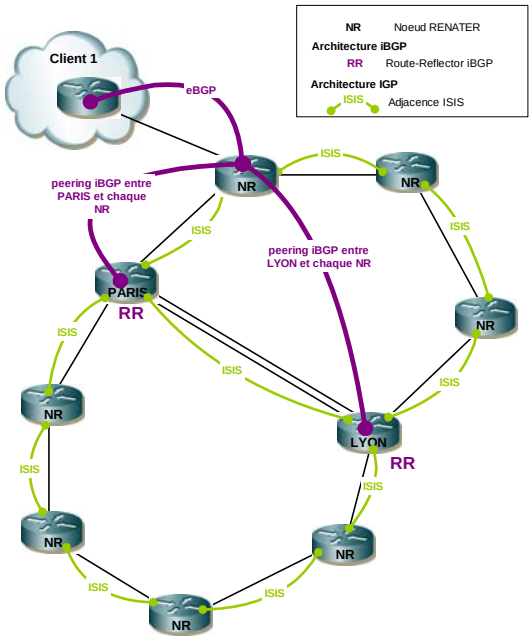


Figure 1 – *Topologie de l'IGP et de l'EGP sur le backbone RENATER*

## 3.3 Mise en oeuvre de MPLS

MPLS a été activé au sein de RENATER-4, afin de proposer des services de VPN MPLS de niveau 2 et 3. Dans la configuration RENATER, tous les routeurs des NR assurent les fonctions PE (« Provider Edge », encapsulation/désencapsulation du label MPLS) et P (« Provider », routage basé sur la lecture des labels).

Le protocole utilisé pour la distribution des labels est LDP (Label Distribution Protocol). Afin d'améliorer la convergence MPLS lors d'un changement d'état d'un lien (coupure/restauration), la fonctionnalité de protection des sessions LDP a été activée. Elle conserve la session LDP entre 2 routeurs adjacents après coupure du lien (sous réserve que le voisin LDP est accessible via un autre chemin) : cela évite de rétablir la session LDP et d'échanger l'ensemble des labels MPLS lors du rétablissement du lien. Par défaut la session LDP est conservée pendant un temps infini.

Le re-calcul d'un lien logique MPLS (ou LSP MPLS) s'appuie sur l'IGP IS-IS.

A l'heure actuelle, le MPLS-TE (MPLS Traffic Engineering) n'est pas appliqué dans le backbone RENATER-4.

# 4 Mise en œuvre de la « fast-convergence » dans le backbone national

Les équipements déployés dans le réseau RENATER-4, sont pour les routeurs des équipements CISCO™ de la famille GSR (124xx) et pour les commutateurs des équipements CISCO™ de la famille Catalyst (essentiellement 45xx et 65xx). La plupart des fonctionnalités mentionnées dans cet article pour l'optimisation des temps de convergence ne sont néanmoins pas spécifiques à ces équipements, mais aux protocoles eux-mêmes : elles peuvent donc être reprises pour d'autres équipements (et donc d'autres constructeurs) dans la mesure où ceux-ci supportent les protocoles en question. Dans le but de faciliter la compréhension de cet article, les paramètres « propriétaires » (et spécifiques aux équipements et aux types de cartes d'interfaces utilisés dans le réseau RENATER) n'y seront, autant que possible, pas explicités.

Les optimisations permettant d'améliorer la convergence portent essentiellement sur l'IGP et l'EGP.

Les valeurs exprimées dans la suite de cet article sont exprimées en secondes (s), millisecondes (ms) et microsecondes (µs).

## 4.1 Optimisations de l'IGP : « fast-convergence » IS-IS

### *4.1.1 Objectifs des optimisations IS-IS*

Avant le début des optimisations IS-IS menées sur le réseau RENATER, le temps de convergence IS-IS estimé était de 1 s à 30 s, suivant le type de liaison.

L'objectif était alors d'obtenir un temps de convergence après optimisation, de quelques centaines de millisecondes et en tous cas bien inférieur à la seconde.

A cet effet, plusieurs axes de recherche ont été explorés, à savoir l'optimisation de la détection d'un changement d'état (ou coupure) de lien par IS-IS, la limitation de la convergence en cas d'instabilités, ainsi que de nombreuses optimisations liées au protocole IS-IS lui-même.

### *4.1.2 Optimisation de la détection d'un changement d'état de lien*

Les optimisations qui peuvent être apportées dépendent du type d'interface. Dans le cas du réseau RENATER, il convient de distinguer les interfaces de type :

- POS (SONET), qui servent à connecter les liaisons 2,5 Gbit/s inter-NR du cœur (cf. figure 2),
- Ethernet (1 Gbit/s et 10 Gbit/s) qui sont plutôt déployées sur l'infrastructure métropolitaine en Ile-de-France (cf. figure 3).

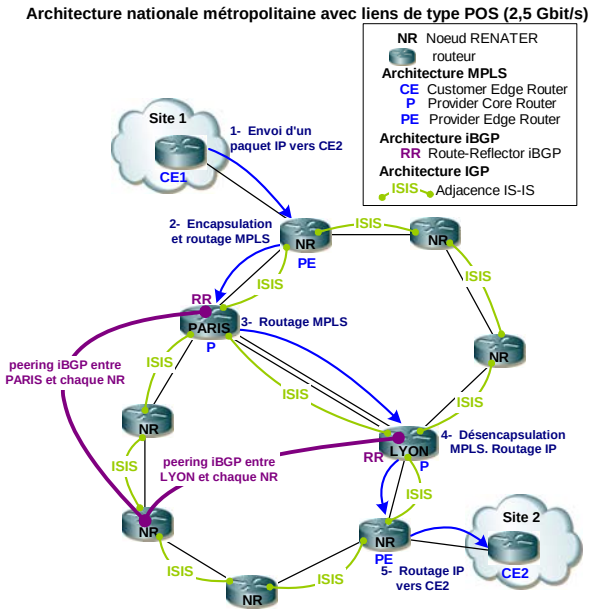


Figure 2 – *topologie de routage sur l'infrastructure nationale métropolitaine du réseau RENATER*

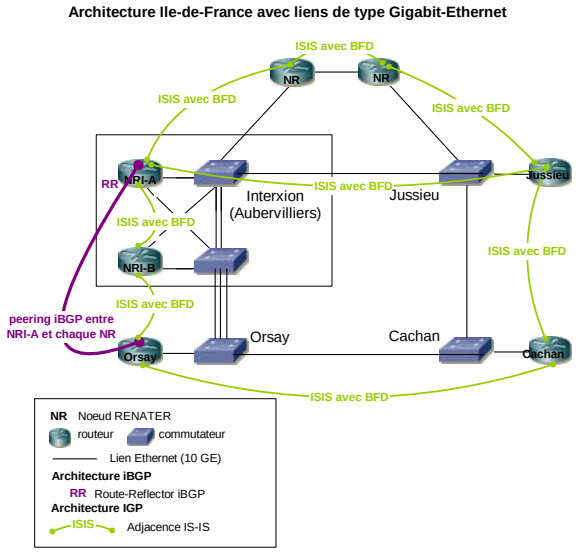


Figure 3 – *topologie de routage sur l'infrastructure RENATER en Ile de France*

### *4.1.3 Optimisations associées aux interfaces POS*

**– Paramètre « carrier delay »**

Ce paramètre détermine le délai entre la prise en compte de la coupure par le système et le passage de l'interface à l'état DOWN. Par défaut, ce délai est de 2 s. Sur une interface POS, il est recommandé de le positionner à 0 s.

**– Paramètre « trigger delay »**

Le « POS trigger delay » correspond au temps de prise en compte par le système d'une alarme SONET/SDH. Dans le cas d'une interface POS protégée, le but est de laisser le temps nécessaire à la réalisation du basculement sur le lien de secours (la valeur du « POS trigger delay » est alors configurée à 60 ms, le secours SDH se faisant en environ 50 ms). Pour une interface POS non protégée la valeur du « POS trigger delay » est à positionner à 0 ms

.

**– « IP Event Dampening »**

L'IP Event Dampening s'active au niveau d'une interface pour pallier les problèmes d'instabilité du réseau liés à des convergences successives en cas de fréquents changements d'état. L'interface se voit attribuer une pénalité à chaque instabilité. Au-delà d'un certain seuil, l'interface n'est plus prise en compte par IS-IS pendant un temps donné.

Il est inutile d'activer l' « IP event Dampening » sur les liens POS car ces derniers possèdent nativement des fonctionnalités de pénalisations et désactivations. En effet, après disparition d'une alarme SDH, l'interface POS attend 10 s avant de passer à l'état UP.

### *4.1.4 Optimisations associées aux interfaces Ethernet*

**– « BFD (Bidirectional Forwarding Detection) »**

BFD optimise le temps de convergence des protocoles de routage (IS-IS, OSPF, BGP). A noter qu'un groupe de travail a été spécifiquement créé à l'IETF à propos de BFD et de ses extensions.

BFD détecte rapidement une perte de connectivité IP entre 2 routeurs adjacents (au niveau logique). Il s'appuie sur l'échange à une fréquence très élevée de paquets HELLO de très petite taille. Au bout d'un certain nombre de HELLO sans acquittement (expiration du « hold-time » BFD), il remonte l'information de perte d'adjacence au(x) protocole(s) de routage pour le(s)quel(s) il a été activé. BFD peut fonctionner avec les protocoles IS-IS, OSPF et BGP (dans un environnement eBGP uniquement, lorsque les peerings BGP sont montés sur les adresses IP d'interconnexion).

L'avantage des paquets HELLO BFD sur les HELLO de IS-IS ou OSPF ou encore les « keep-alive » BGP est qu'ils nécessitent peu de traitement par le routeur et sont ainsi moins consommateurs de CPU. Ils peuvent donc être émis avec une périodicité plus courte, de l'ordre de la cinquantaine de millisecondes.

A réception de la notification BFD de perte de connectivité, le protocole de routage converge pour mettre sa table à jour.

**– BFD pour IS-IS**

Cette fonctionnalité réduit à 150 ms le temps de détection par IS-IS d'une coupure de lien (à l'exception des liens POS pour lesquels le temps de détection est inférieur à 50 ms).

BFD présente tout son intérêt sur des liaisons Ethernet transitant par des commutateurs. BFD est déployé sur tous les liens Ethernet (liens Gigabit-Ethernet - 1 Gbit/s et 10 Gbit/s) du réseau RENATER, ainsi que sur les interfaces

Gigabit-Ethernet interconnectant les commutateurs aux routeurs au sein des NR.

Pour la détection d'une coupure de lien, le choix de BFD a été préféré à une solution de type IS-IS « Fast-HELLO » qui aurait alors consisté à augmenter la fréquence d'envoi des paquets HELLO pour détecter une perte d'adjacence IS-IS en 1 s (la valeur par défaut étant 30 s). En effet, l'échange de paquets BFD se fait avec une périodicité plus courte avec tous les voisins (périodicité maximum fixée à 50 ms), et si un paquet BFD n'est pas reçu pendant un intervalle de temps déterminé (valeur minimale fixée à 150 ms), la perte de connectivité est signalée et remontée à IS-IS, ce qui en fait une solution beaucoup plus performante que le « Fast-HELLO ».

**– Paramètre « carrier delay »**

Ce paramètre est positionné à 0 s sur les interfaces Ethernet du backbone RENATER. A noter qu'il est intéressant de réduire la valeur de ce paramètre au maximum surtout lorsque BFD n'est pas disponible sur les interfaces.

**– « IP Event Dampening »**

La fonctionnalité est activée sur toutes les interfaces Ethernet du backbone, pour réduire les effets de bord en cas d'instabilité sur une liaison.

**– Mode IS-IS point-à-point**

Sur une liaison point-à-point, le processus IS-IS n'a qu'une seule adjacence à établir entre les 2 routeurs interconnectés. En revanche, dans un environnement point-à-multipoint de type LAN, la mise en place d'IS-IS est plus complexe et nécessite l'activation de mécanismes spécifiques plus consommateurs en ressources.

Ces mécanismes IS-IS propres à un environnement LAN sont activés par défaut lors de la mise en place de IS-IS sur une interface Ethernet mais peuvent être inhibés. IS-IS doit donc être informé explicitement que les interfaces Ethernet du backbone RENATER sont des liaisons point-à-point et non pas LAN.

### *4.1.5 Optimisations du protocole IS-IS*

**– Fonctionnalité « Incremental SPF »**

Par défaut, chaque routeur recalcule le chemin optimal pour toutes les routes référencées dans l'IGP en cas de perte/rétablissement d'une adjacence IS-IS. L' « Incremental SPF (Shortest Path First) » effectue un re-calcul restreint aux routes IS-IS impactées par le changement.

Cette fonctionnalité n'a pas été retenue car elle apporte peu de gain. En effet, le temps de calcul SPF est négligeable comparé au temps nécessaire pour mettre la table de routage (RIB - Routing Information Base) à jour. Ainsi pour le calcul SPF il faut compter 40 µs par nœud, ce qui, pour le réseau RENATER représente un total de seulement 2 ms (l'aire IS-IS du réseau RENATER étant constituée d'environ 50 routeurs), à comparer à la mise à jour de la RIB qui prend environ 150 ms pour 100 préfixes (soit un total de 600 ms pour 400 préfixes dans le cas du réseau RENATER).

**– « LSP flooding »**

Lorqu'un routeur détecte un changement d'état d'une de ses interfaces (coupure/rétablissement d'un lien), il génère un paquet LSP IS-IS (LSP : Link State PDU – Protocol Data Unit) et l'envoie à ses voisins IS-IS qui le retransmettent à leur tour à tous leurs voisins IS-IS. Le LSP IS-IS est ainsi propagé sur tout le réseau. Ces LSP IS-IS sont les événements déclencheurs du re-calcul SPF par les nœuds participant à l'IGP.

Par défaut, un routeur attend 33 ms entre l'envoi de 2 LSP IS-IS consécutifs qu'il reçoit de ses adjacences IS-IS. Avec le « LSP fast-flooding », dès que le routeur reçoit un LSP IS-IS, il le transmet immédiatement à toutes ses adjacences IS-IS.

Le nombre maximum d'adjacences IS-IS que peut avoir un nœud du backbone dans le cadre du « fast-flooding » a été limité à 15 (recommandation CISCO™) sur les équipements concernés.

**– « Back-off timers »**

Le principe des « backoff timers » est de réagir rapidement à un événement isolé comme une coupure de lien, et plus lentement lorsque un nombre important de LSP IS-IS sont propagés sur le réseau dans un délai très court (dans le cas de la perte d'un nœud pas exemple, qui a pour conséquence la perte de plusieurs liaisons à la fois).

A noter, qu'en fonction des valeurs configurées pour ces « timers », on peut optimiser le re-calcul SPF soit pour traiter une perte de lien ou une perte de nœud (coupure d'un NR). Dans le cas du réseau RENATER, la priorité a été donnée pour le traitement d'une coupure d'un lien (dans la mesure où les NR sont desservis dans leur quasi-totalité par au moins 2 liaisons WAN), plus que l'isolement d'un NR (rarissime). Dans la pratique, le re-calcul SPF (partiel ou complet) est déclenché seulement 1 ms après la réception du premier LSP IS-IS, or une perte de nœud entraîne la coupure de plusieurs liens du backbone et donc la génération de plusieurs LSP IS-IS. Le calcul SPF et la mise à jour de la RIB seront donc lancés avant réception de tous les LSP relatifs au même incident. Cette convergence, qui dure au plus 600 ms sur RENATER, aboutit à une table de routage erronée. Un nouveau calcul SPF est relancé 10 ms après le premier, pour prendre en compte les LSP IS-IS non considérés lors du premier calcul.

**– Priorisation de préfixes IS-IS lors de la mise à jour de la table de routage**

La table de routage (RIB) est mise à jour après réalisation du calcul SPF. Il est possible de prioriser certains préfixes IS-IS pour qu'ils soient mis à jour dans la RIB avant les préfixes IS-IS de plus faible priorité. Le temps de

convergence est donc réduit pour les préfixes IS-IS importants.

Certains préfixes IS-IS « stratégiques », notamment ceux associés aux interconnexions avec des partenaires internationaux (GEANT, SFINX et accès de transit IP) sont ainsi priorisés : le temps de calcul SPF pour ces préfixes prioritaires passe ainsi de plusieurs centaines de ms (jusqu'à 600 ms) à environ 6 ms.

### *4.1.6 Bilan des optimisations relatives à la convergence de l'IGP*

| Evénement | Détection de la perte de connectivité | Délai de retransmission des LSP IS-IS relatifs à l'événement | Temps d'exécution du calcul SPF | Temps de mise à jour de la RIB (table de routage IS-IS) |
|---|---|---|---|---|
| **Coupure d'un lien de type POS** | SANS FC (Fast Convergence) ~ 2,05 s<br>AVEC FC < 20 ms | SANS FC ~ 33 ms<br>AVEC FC ~ 0 ms | | |
| **Coupure d'un lien Ethernet point-à-point direct entre 2 routeurs (pas de commutateur intermédiaire)** | SANS FC ~ 2,3 s<br>AVEC FC ~ 150 ms | SANS FC ~ 33 ms<br>AVEC FC ~ 0 ms | | SANS / AVEC FC ~ 600ms |
| **Coupure d'un lien Ethernet transitant par un commutateur** | SANS FC ~ 30 s<br>AVEC FC ~ 150 ms | SANS FC ~ 33 ms<br>AVEC FC ~ 0 ms | SANS / AVEC FC ~ 2ms | **Mise à jour des accès vers l'international :** SANS FC ~ 600 ms<br>AVEC FC ~ 6 ms |
| **Perte d'un routeur RENATER (exemple d'un routeur ayant 5 adjacences ISIS)** | SANS FC ~ 30 s<br>AVEC FC ~ 150 ms | SANS FC ~ 165 ms<br>AVEC FC ~ 0 ms | | |

**REMARQUE :** Le calcul SPF et la mise à jour de la RIB peuvent être réalisés plusieurs fois pour un même événement. Les données fournies correspondent au temps de réalisation d'un unique calcul SPF ou mise à jour de la RIB.

Tableau 1 – *Bilan des optimisations relatives à la convergence de l'IGP*

## 4.2 Optimisations globales apportées pour l'EGP : « fast-convergence » BGP

### *4.2.1 Mode de raccordement sur un NR*

Les modes de raccordements possibles sur un NR sont les suivants (indépendamment du fait que le partenaire dispose d'un simple ou d'un double attachement) :

– Mode 1 (cf. figure 4) : raccordement du partenaire sur le commutateur du NR, via un port Ethernet (port 10/100M, ou encore 1 ou 10 Gigabit-Ethernet). Ce mode de raccordement est le plus courant (le commutateur étant raccordé au routeur via une ou plusieurs interfaces 1 ou 10 Gigabit-Ethernet) ;

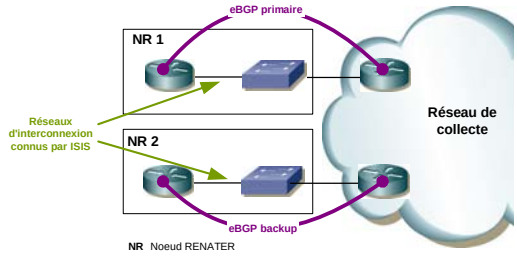


Figure 4 – *mode de raccordement sur un NR (mode 1)*

– Mode 2 (cf. figure 5) : raccordement du partenaire sur le routeur du NR, via un port Gigabit-Ethernet.

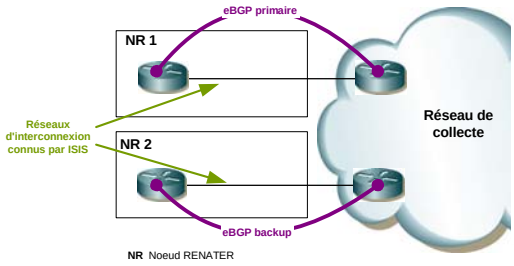


Figure 5 – *mode de raccordement sur un NR (mode 2)*

Les fonctionnalités mises en jeu pour assurer une convergence rapide du réseau sont différentes selon que la coupure est directement visible par le routeur du NR (passage de l'interface du routeur à l'état « DOWN ») ou qu'elle intervient derrière le commutateur du NR (l'interface du routeur reste dans l'état « UP »).

La « fast-convergence » et les optimisations associées prennent bien entendu une autre dimension dans le cas d'un « dual-homing », mais il est aussi possible d'apporter des optimisations dans le cas d'un partenaire raccordé via un simple attachement sur un NR. A ce propos, il est important de noter que les exemples de réseaux de collecte en double attachement sur un NR (ou encore raccordés à 2 NR distincts) tendent à se généraliser, et à ce titre de tels réseaux déploient une infrastructure redondante (doublement du routeur de « peering » avec RENATER ainsi que des équipements de concentration pour les équipements de cœur de réseau quand ceux-ci sont distincts des routeurs de « peering »).

Le cas d'un double attachement dans un premier temps, et d'un simple attachement dans un second temps, seront donc traités ci-après.

Le mode 1 sera particulièrement détaillé car le comportement du routeur dans le cas du mode 2 est assimilable à une coupure de lien entre le routeur et le commutateur du NR dans le mode 1.

### *4.2.2 Objectifs des optimisations BGP*

Afin d'améliorer la convergence BGP, plusieurs axes de recherche ont été explorés, à savoir réduire la perte de paquets BGP, limiter les « flaps » (instabilités) de préfixes, traiter plus rapidement les paquets BGP, réduire le délai d'annonce des préfixes et accélérer les mises à jour des tables de routage, améliorer le temps de détection par BGP d'un changement de topologie.

### *4.2.3 Optimisations iBGP*

Il n'a pas été possible d'activer de mécanisme propre à iBGP permettant d'améliorer les délais d'échanges des préfixes au sein d'un AS.

Les améliorations possibles au niveau de iBGP sont plus génériques et s'appliquent également à un environnement eBGP. Ainsi, une optimisation commune possible concerne l'intervalle minimum entre 2 notifications de mise à jour de la table BGP : ce paramètre a été configuré à 0 s pour l'ensemble des routeurs RENATER pour les peerings iBGP, et peut être généralisé au niveau des peerings eBGP.

### *4.2.4 Optimisations eBGP*

Il est important de faire ici un inventaire des principales fonctionnalités disponibles pour les optimisations eBGP, lesquelles seront ensuite précisées pour le cas d'un réseau de collecte en double attachement sur le réseau RENATER par rapport à un simple attachement.

Cette section est importante dans la mesure où elle précise notamment ce qui doit être activé sur les routeurs (celui du NR, comme celui du réseau de collecte), et à quel niveau cela doit l'être (fonctionnalité globale à l'équipement, ou à activer sur une interface, ou encore sur un peering).

**– Fonctionnalité « fast external fallover »**

L'objet de cette fonctionnalité est de fermer la session eBGP automatiquement dès la détection de la coupure physique du lien.

Pour être efficace, elle devra donc être activée sur le routeur du NR qui a établi le peering eBGP avec le réseau de collecte, et réciproquement sur le routeur du réseau de collecte. Néanmoins, comme on le verra plus loin dans l'article, il peut exister des cas où il est intéressant de la désactiver.

Cette fonctionnalité est activée par défaut.

**– Fonctionnalité « next-hop tracking »**

La perte de connectivité est détectée par l'IGP, qui prévient le processus BGP d'un changement de topologie.

Cette fonctionnalité est activée par défaut. Elle est appliquée de manière globale sur le routeur du NR.

Elle devra être activée sur le routeur du réseau de collecte seulement si ce dernier s'appuie également sur un IGP (IS-IS ou OSPF) pour son cœur de réseau régional ou métropolitain.

Le paramètre « next-hop trigger delay » détermine le temps que va attendre le processus BGP après réception d'une notification « next-hop tracking » pour mettre à jour les entrées BGP de la RIB : sa valeur par défaut est de 5 s. Cette deuxième solution permet théoriquement une convergence aussi rapide que le « fast external fallover » si le paramètre « next-hop trigger delay » est positionné à 0 s.

Le « next-hop tracking » a l'avantage de ne pas faire tomber la session BGP.

**– « IP event dampening »**

Comme mentionné en § 4.1.4, la fonctionnalité « IP event dampening » est activée sur toutes les interfaces Ethernet du réseau RENATER.

Cette fonctionnalité devra donc aussi être activée sur l'interface Ethernet du routeur RENATER (vers le commutateur du NR) et, pour être synchrone, sur l'interface Ethernet du routeur du réseau de collecte (vers le commutateur du NR). L' « IP event dampening » permettra d'éviter une convergence inutile (ouverture/fermeture successives des peerings BGP) en cas de bagot. Elle sera d'autant plus utile si la fonctionnalité de « fast external fallover » est activée (cette dernière fermant automatiquement la session eBGP en cas de coupure physique du lien).

**– Paramètre « carrier delay »**

Nous avons vu que le paramètre « carrier delay » était utilisé pour les liens Ethernet du réseau RENATER (cf. § 4.1.4) : il détermine le délai entre la détection d'une perte de lien et le passage de l'interface à l'état « DOWN ».

Cette fonctionnalité devra donc aussi être activée sur l'interface Ethernet du routeur RENATER (vers le commutateur du NR) et, pour être synchrone, sur l'interface Ethernet du routeur du réseau de collecte (vers le commutateur du NR).

La valeur par défaut est du paramètre « carrier delay » est de 2 s.

**– Utilisation de BFD**

Les cas d'utilisation de BFD pour BGP seront détaillés ci-après.

## 4.3 Cas d'un réseau de collecte en double attachement sur le réseau RENATER

### *4.3.1 Coupure du lien entre le routeur et le commutateur du NR*

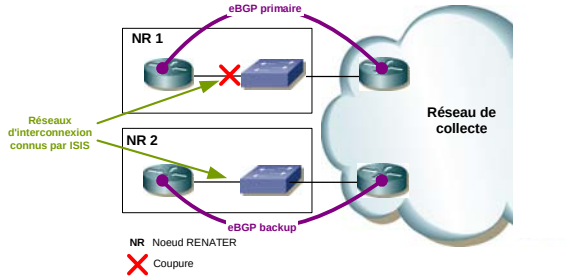


Figure 6 – *Coupure de lien entre le routeur et le commutateur du NR (mode 1)*

La convergence sera optimisée grâce aux fonctionnalités de « fast external fallover » ou de « next-hop tracking » ainsi que de « carrier delay » (cf. optimisations listées en § 4.2.4).

**– Paramètre « carrier delay »**

La valeur de ce paramètre est à positionner à 500 ms sur l'interface vers le réseau de collecte (la valeur par défaut étant de 2 s). L'intérêt de mettre une valeur différente de 0 s est de masquer un « flap » très court lorsque le « fast external fallover » est activé, pour ne pas fermer trop rapidement la session BGP.

Le passage de l'interface de raccordement avec le réseau de collecte à l'état DOWN et la fermeture de la session BGP auront donc lieu 750 ms après la coupure : 250 ms (valeur constatée) nécessaires au système pour détecter la coupure sur un lien Ethernet + 500 ms associées au « carrier delay »).

**– Autres optimisations**

Outre les optimisations BGP mentionnées précédemment, celles indiquées pour l'IS-IS dans la rubrique § 4.1.4 (optimisations IS-IS sur des liens de type Ethernet) sont bien sûr appliquées au niveau de l'interconnexion du routeur et du commutateur RENATER.

### *4.3.2 Coupure du lien entre le commutateur du NR et le routeur du réseau de collecte*

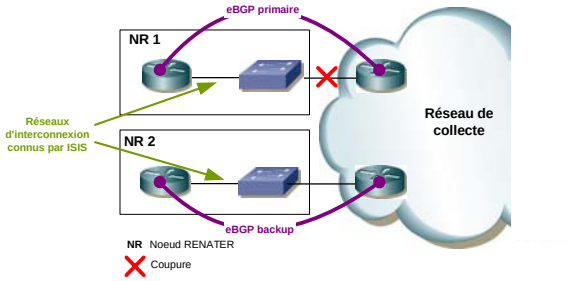


Figure 7 – *Coupure de lien entre le commutateur du NR et le routeur du réseau de collecte (mode 1)*

Dans ce cas, la perte de connectivité physique n'est pas visible par le routeur du NR (puisque un commutateur se trouve intercalé entre le routeur du réseau de collecte et celui de RENATER). Les fonctionnalités mises en jeu dans le cas précédent ne s'appliquent donc pas (cf. § 4.3.1, cas de la coupure du lien entre le routeur et le commutateur du NR).

**– Utilisation de BFD**

La solution adaptée à ce type de coupure est BFD pour BGP. Cette fonctionnalité permet de réduire à 750 ms le temps de détection d'une perte de connectivité. Si le routeur du NR ne reçoit pas de paquet BFD de son voisin BGP pendant plus de 750 ms, il ferme automatiquement la session BGP associée.

BFD est configurable avec un « hold-time » d'une valeur minimale de 150 ms. Sur RENATER, la valeur de 750 ms a été choisie pour éviter des fermetures de sessions BGP trop rapides en cas de « flap » de lien.

En positionnant BFD à 750 ms, nous aurons les mêmes temps de réaction, que la coupure intervienne sur le lien interconnectant le routeur et le commutateur du NR ou sur le lien interconnectant le commutateur au routeur du réseau de collecte.

### *4.3.3 Perte du routeur du NR*

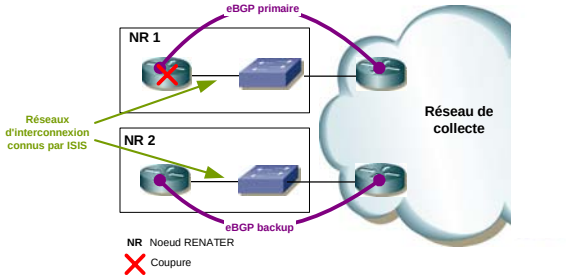


Figure 8 – *Coupure du routeur du NR (mode 1)*

Dans ce cas, la fonctionnalité de « next-hop tracking » (perte de connectivité détectée par l'IGP) est utile, et la valeur par défaut de 5 s du « next-hop trigger delay » peut

être ramenée à 1 s, pour augmenter la réactivité. Ce délai de 1 s est une sécurité pour s'assurer que l'IGP a terminé sa convergence et que la RIB soit mise à jour. Sans cette garantie, le « scan » de la table BGP, lancé par le processus BGP et qui se base sur les entrées IGP de la RIB, aboutirait à une convergence BGP erronée.

## 4.4 Cas d'un réseau de collecte en simple attachement sur le réseau RENATER

Cette section décrit les fonctionnalités qui peuvent être utilisées dans le cas de réseaux de collecte qui sont en simple attachement sur le NR, et pour lesquels cela constitue la seule porte de sortie pour leur accès Internet.

### *4.4.1 Coupure du lien entre le routeur et le commutateur du NR*

L'optimisation de la convergence dans le cas de cette coupure sera traitée par l'utilisation des fonctionnalités de « next-hop tracking », « IP event dampening » et « carrier delay ».

**– Fonctionnalité « fast external fallover »**

Dans le cas d'un réseau de collecte en simple attachement sur le NR, l'objectif est de garder autant que possible la session BGP active.

Le « fast external fallover » décrit en § 4.2.4 n'est donc ici pas forcément un avantage, dans la mesure où cela revient à fermer très rapidement la session BGP (et donc à couper la connectivité IP). Il faut alors désactiver cette fonctionnalité sur le routeur du NR, mais aussi au niveau du routeur du réseau de collecte. La session BGP sera alors maintenue jusqu'à expiration du BGP « keepalive hold-time » (180 s sur un routeur CISCO, mais différente chez d'autres constructeurs).

**– Fonctionnalité « next-hop tracking »**

La convergence BGP, pour les préfixes annoncés par les réseaux de collecte, sera déclenchée par le « next-hop tracking ». Elle aura donc lieu avant la fermeture de la session BGP (puisque le « fast external fallover » est désactivé).

**– Paramètre « carrier delay »**

Ce paramètre peut être abaissé à 500 ms.

### *4.4.2 Coupure du lien entre le commutateur du NR et le routeur du réseau de collecte*

Dans le cas d'un réseau de collecte en simple attachement sur le NR, l'objectif est de garder autant que possible la session BGP active. A cet effet BFD ne sera pas configuré sur l'interconnexion physique entre le commutateur du NR et le routeur du réseau de collecte. Dans ce cas, la convergence BGP aura lieu seulement au moment de la fermeture de la session BGP, à savoir l'expiration du BGP « keepalive hold-time » (tel que mentionné en § 4.4.1).

La fonctionnalité d' « IP event dampening » contribuera à maintenir la session BGP active sans qu'elle ne soit pénalisée par le moindre bagot.

### *4.4.3 Bilan des optimisations relatives à la convergence de l'EGP*

| **Evénement** | | **Détection de la perte de connectivité par BGP** |
|---|---|---|
| **Coupure de lien vers un site ou un réseau de collecte en double attachement sur RENATER** | ***Coupure du lien entre le routeur et le commutateur du NR*** | SANS FC (Fast Convergence)<br>~ 2,3 s<br>AVEC FC<br>~ 750 ms |
| | ***Coupure du lien entre le commutateur du NR et le routeur du réseau de collecte*** | SANS FC<br>~ 180 s (CISCO)<br>AVEC FC<br>~ 750 ms |
| **Coupure de lien vers un site ou un réseau de collecte en simple attachement sur RENATER** | ***Coupure du lien entre le routeur et le commutateur du NR*** | SANS FC<br>~ 2,3 s<br>(avec fermeture de la session BGP)<br><br>AVEC FC<br>~ 1,75 s<br>(avec conservation de la session BGP) |
| | ***Coupure du lien entre le commutateur du NR et le routeur du réseau de collecte*** | SANS / AVEC FC<br>~ 180 s (CISCO) |
| **Perte du routeur du NR raccordant le réseau de collecte** | | SANS FC<br>~ LoC(IGP) + 5 s<br>AVEC FC<br>~ LoC(IGP) + 1 s |
| **Rupture d'un des accès IP transit (Paris / Lyon) de RENATER** | | SANS FC<br>~ 2,05 s<br>AVEC FC<br>~ 520 ms |
| **Rupture de l'accès IP primaire GEANT** | | SANS FC<br>~ 90 s<br>AVEC FC<br>~ 1 s |

**REMARQUE** : LoC(IGP) = Délai de convergence de l'IGP. Pour obtenir une estimation de cette valeur, se reporter au cas de la perte d'un routeur RENATER dans le tableau de synthèse illustrant les temps de convergence IS-IS.

Tableau 2 – *Bilan des optimisations relatives à la convergence de l'EGP*

# 5 Conclusion

La mise en place de la « fast-convergence » dans le réseau RENATER a nécessité une étude approfondie de l'ensemble des protocoles utilisés dans le réseau (IS-IS, BGP, MPLS).

Les optimisations mises en œuvre améliorent significativement les temps de convergence dans le réseau, le rendant beaucoup plus « transparent » aux applications interactives ou temps réel. D'abord appliquées dans le cœur de réseau, puis en périphérie sur les interconnexions avec les « éléments extérieurs » en double attachement (GEANT-2, SFINX™, accès IP transit), il reste désormais à les adapter et intégrer aux réseaux de collecte et réseaux de campus.

# Annexes

### Glossaire

AS : Autonomous System
BFD : Bidirectional Forwarding Detection
BGP : Border Gateway Protocol
CPU : Central Processing Unit
eBGP : External Border Gateway Protocol
EGP : Exterior Gateway Protocol
GIP : Groupement d'Intérêt Public
GPS : Global Positioning System
iBGP : Internal Border Gateway Protocol
IETF : Internet Engineering Task Force
IGP : Interior Gateway Protocol
IP : Internet Protocol
IS-IS : Intermediate System to Intermediate System
LAN : Local Area Network
LDP : Label Distribution Protocol
LSP MPLS : Label Switched Path MPLS
LSP IS-IS : Link State PDU IS-IS
MPLS : Multi-Protocol Label Switching
MPLS-TE : MPLS Traffic Engineering
MSS : Maximum Segment Size
MTU : Maximum Transfer Unit
NOC : Network Operations Centre
NR : Nœud RENATER
OSPF : Open Shortest Path First
PDU : Protocol Data Unit
POS : Packet Over SONET
RENATER : Réseau National pour la Technologie l'Enseignement et la Recherche
RIB : Routing Information Base
SDH : Synchronous Digital Hierarchy
SFINX : Service for French Internet eXchange
SONET : Synchronous Optical NETwork
SPF : Shortest Path First
TCP : Transfer Control Protocol
ToIP : Telephony over IP
WAN : Wide Area Network