# Fiabilisation d'une architecture DNS

Philippe Pegon
Centre Réseau Communication, Université Louis Pasteur
7, rue René Descartes
67084 Strasbourg
Philippe.Pegon@crc.u-strasbg.fr
Date : 14 octobre 2003

### Résumé

*Les utilisateurs du réseau sont de plus en plus exigeants sur sa disponibilité. L'objectif politique assigné au CRC, l'opérateur du réseau Osiris de l'enseignement supérieur et de la recherche à Strasbourg, est d'atteindre une disponibilité de 99,9 %. Dans ce cadre, l'architecture DNS d'Osiris a été refondue au cours du premier semestre 2003.*
*Cette nouvelle architecture, adaptée à une topologie réseau robuste, a pour but de diminuer le nombre de serveurs DNS délocalisés, au profit d'un serveur rendu plus fiable.*
*Cette contribution présente la réflexion menée sur les problèmes de fiabilité du DNS, la nouvelle architecture mise en place, le protocole assurant la redondance en moins de 4 secondes, ainsi que les aspects concrets de la mise en œuvre pour ceux qui voudraient appliquer l'idée dans d'autres contextes (campus, laboratoires, etc.). Le recul de quelques mois d'exploitation permet de dresser le bilan des avantages et des inconvénients.*

### Mots clefs

DNS, VRRP, redondance, disponibilité

## 1 Introduction

La plupart des nouveaux arrivants sur le réseau n'en ont pas connu les débuts, parfois chaotiques, et montrent une certaine incompréhension lorsque des techniciens expliquent que l'informatique n'est pas comme l'électricité, qu'il ne suffit pas d'appuyer sur l'interrupteur pour que cela marche, et qu'il peut y avoir des bugs ou des problèmes parfois inexplicables. L'exigence de disponibilité croît avec la dépendance vis-à-vis des outils informatiques, eux-mêmes de plus en plus dépendants du réseau.
Le Centre Réseau Communication (CRC) de l'Université Louis Pasteur gère et exploite le réseau métropolitain Osiris pour le compte des établissements d'enseignement supérieur et de recherche strasbourgeois. Les instances politiques ont défini un objectif fort en termes de disponibilité : 99,9 %, soit moins de 4 heures d'interruption par an au maximum.

Pour attteindre cet objectif ambitieux, le CRC a initié une démarche d'ampleur, identifiée sous le nom de «projet Osiris 2 », et déclinée en sous-projets :
- redondance de l'infrastructure optique métropolitaine ;
- sécurisation électrique des nœuds de dorsale, et de la centaine de sites connectés ;
- climatisation ou rafraîchissement des locaux et armoires réseau ;
- renouvellement de l'ensemble du matériel actif (nœuds de dorsale et sites connectés) ;
- renouvellement des serveurs ;
- évolution de l'architecture des services, et plus particulièrement le volet « DNS », le CRC gérant les serveurs DNS pour l'ensemble d'Osiris.

Le reste de cet article présente les motivations de l'évolution de l'architecture DNS, l'ancienne architecture et ses limitations, la nouvelle architecture et sa mise en œuvre concrète dans l'environnement Osiris.

## 2 Pourquoi faire évoluer l'architecture DNS ?

### 2.1 Motivation

Tous les efforts placés dans la fiabilisation du réseau peuvent être réduits à néant par la panne d'un serveur de noms. Le réseau a beau continuer à router des datagrammes IP, si plus aucun utilisateur ne peut initier une nouvelle connexion (ouverture de page Web ou autre), c'est l'émeute ! Pour l'utilisateur, le résultat d'une panne de serveur DNS ou de la coupure d'une fibre optique est le même : « le réseau est encore en panne ». Il faut dire que les messages d'erreur des applications ne les aident pas toujours à faire la distinction...

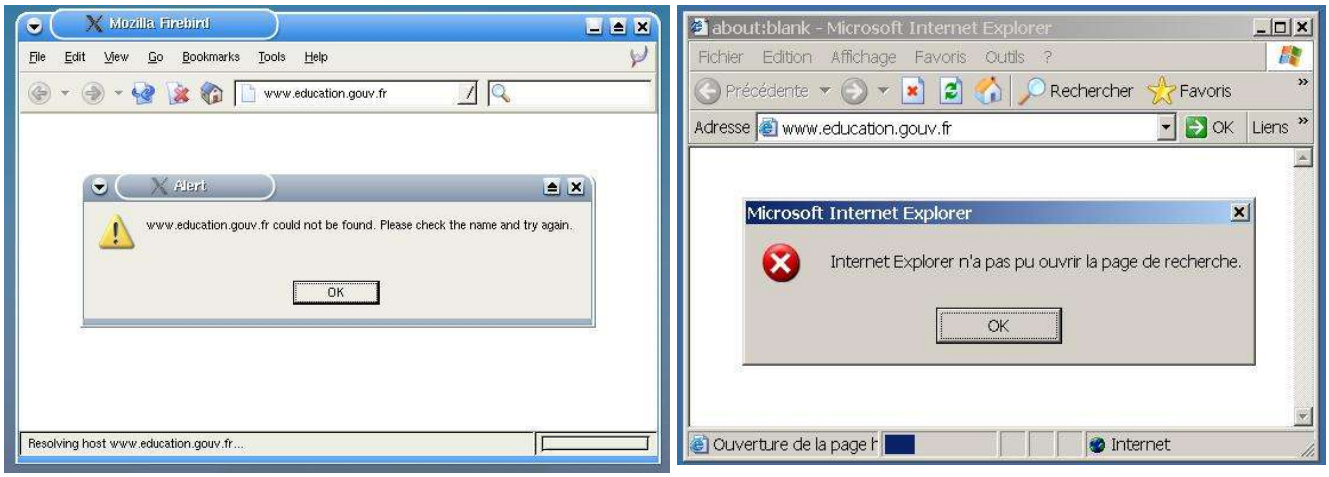


Figure 1 – *Le comportement de différents navigateurs face à un problème DNS*

Assurer une disponibilité réseau implique donc d'assurer une disponibilité du service DNS au moins aussi bonne.

## 2.2 L'architecture des années 90

Le réseau métropolitain Osiris fédère plus de 18 000 ordinateurs, répartis sur une centaine de bâtiments regroupés principalement sur quatre grands campus : Cronenbourg, Illkirch-Graffenstaden, Médecine et l'Esplanade.

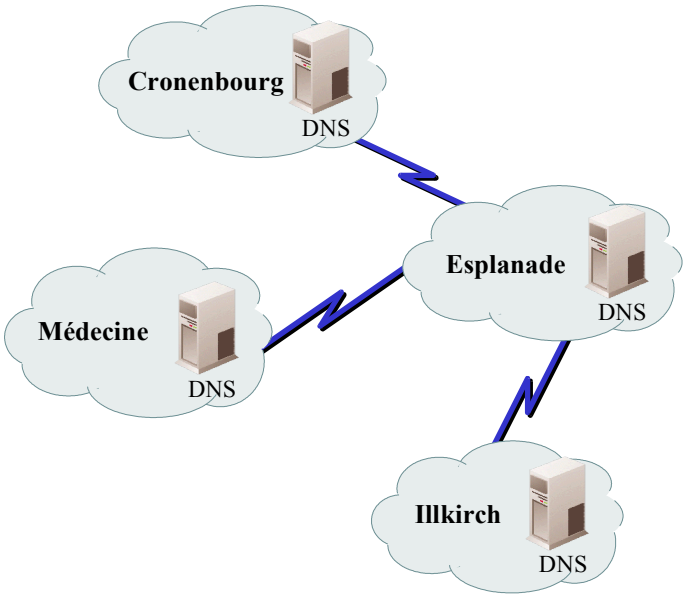


Figure 2 – *Osiris dans les années 90*

Le réseau Osiris, comme la plupart des réseaux, était initialement bâti sur une infrastructure de liaisons spécialisées à bas débit. L'architecture des serveurs DNS s'est naturellement calquée sur ce modèle et ses contraintes : pour des raisons d'efficacité, les serveurs de noms ont été répartis sur chacun de ces sites de manière à minimiser le trafic inter-campus (grâce au cache local des serveurs) et à maximiser les performances (grâce à la proximité du serveur répondant aux requêtes des résolveurs de l'ensemble des stations du campus).
Cette architecture bénéficie de la notion de proximité géographique, mais souffre d'une faiblesse évidente : un problème sur un des serveurs DNS affecte l'ensemble des machines du campus concerné. La possibilité de mettre deux serveurs DNS différents dans la configuration du résolveur (`/etc/resolv.conf` sur les machines Unix) n'offre pas de vraie solution, car le deuxième serveur n'est interrogé qu'après expiration du délai pour le premier, d'où une perception de lenteur très pénalisante.
La deuxième faiblesse inhérente à cette architecture est la présence de plusieurs serveurs délocalisés, ce qui entraîne une charge d'administration importante lors des mises à jour de système, de failles de sécurité, etc.

## 2.3 L'architecture du XXIe siècle

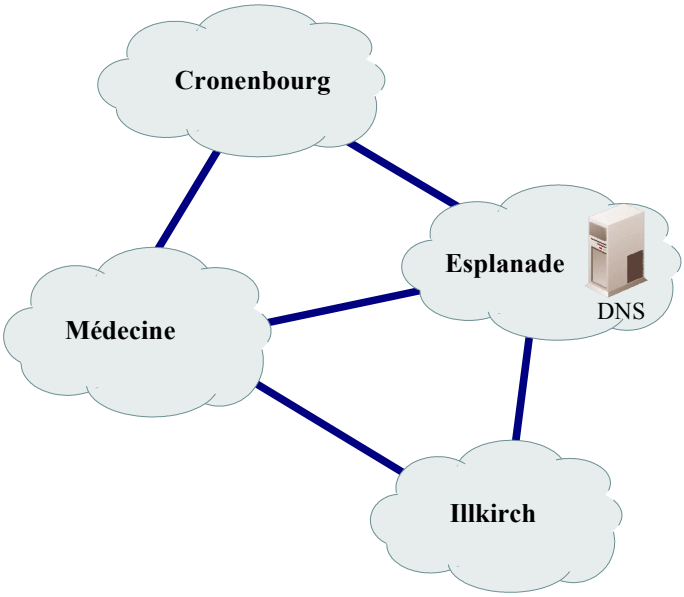


Figure 3 – *Osiris au XXIe siècle*

Si lors de la mise en place initiale du réseau, la notion de proximité géographique avait un sens, comme nous l'avons vu, tous les efforts pour créer un réseau métropolitain à haut débit visent à éliminer les distances ; de ce fait, l'évolution des débits annihile les optimisations liées à la proximité géographique.

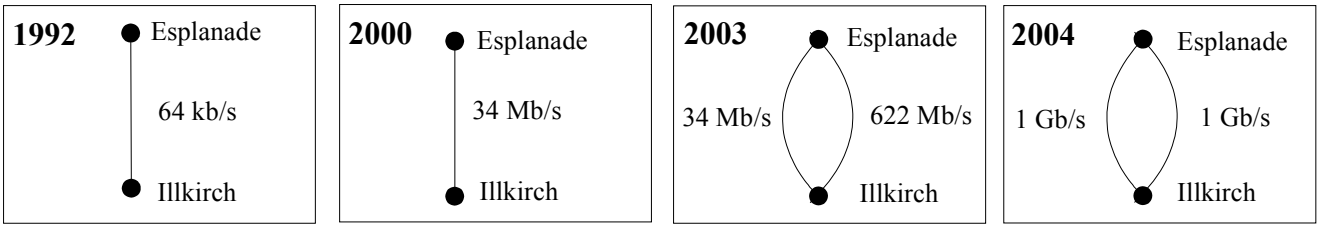


Figure 4 – *Osiris : évolution des débits*

En plus de cette évolution des débits, la mise en place d'une redondance de l'infrastructure optique rend désormais les liaisons inter-campus aussi fiables, sinon plus, qu'un réseau local.

Avec un tel réseau, la notion de proximité géographique n'ayant plus de sens, l'évolution vers un service DNS plus robuste passe par une redéfinition de l'architecture.

# 3 Principe de la nouvelle architecture DNS

Ayant constaté les limitations de l'ancienne architecture, la réflexion s'est rapidement engagée sur l'idée de **concentrer les efforts sur un petit nombre de serveurs centraux**, plutôt que disperser les efforts sur des serveurs délocalisés sur tous les campus.

De plus, compte-tenu du comportement des résolveurs, qui essayent le premier serveur, puis le deuxième serveur, et ainsi de suite jusqu'à la fin de la liste, il importe que la première adresse réponde toujours. De cette constatation est née l'idée de faire **partager une et une seule adresse IP** par notre petit nombre de serveurs centraux, ce qui présente également l'avantage de simplifier la configuration des postes clients, une seule adresse IP étant à retenir pour l'ensemble de la communauté.

La première hypothèse étudiée consistait à utiliser les mécanismes de routage afin que l'adresse IP partagée soit routée vers un des serveurs, soit de façon « maître-esclave » (un seul serveur est actif à un moment donné), soit de façon « symétrique » (les deux serveurs sont actifs et se partagent la charge). Le protocole IPv6 dispose de cette fonctionnalité sous le nom d'adresse « anycast ». Cette solution a été écartée car elle faisait participer les serveurs directement au processus de routage OSPF : les simulations ont montré que les problèmes d'interopérabilité auraient fragilisé l'ensemble du réseau.

La deuxième hypothèse envisagée s'est avérée être la bonne : le protocole VRRP (Virtual Router Redundancy Protocol) [1], initialement prévu comme son nom l'indique pour assurer la redondance de routeurs, peut très bien s'appliquer à un ensemble de serveurs. L'intérêt de cette solution est sa rapidité (reprise en 3 secondes) et son indépendance vis-à-vis des protocoles de propagation de tables de routage, les routeurs (ou serveurs dans notre cas) étant sur le même réseau « local » c'est à dire un VLAN transversal dans le contexte d'Osiris.

Bien évidemment, cette nouvelle architecture a pour conséquence de faire disparaître les serveurs délocalisés ; elle entraîne donc une reconfiguration d'une grande partie des résolveurs des stations clientes. C'est pourquoi un délai de plus d'un an est prévu pour terminer cette transition avant l'extinction des anciens serveurs.

# 4 Fonctionnement de VRRP

Le protocole VRRP fonctionne suivant un principe « maître/esclave » ; il permet à plusieurs routeurs (256 au maximum) situés sur un même réseau local de se secourir mutuellement. Les explications ci-après illustrent le fonctionnement sur deux équipements, le principe étant généralisable.

Les deux machines partagent une même adresse IP, dite adresse IP virtuelle, et également une même adresse MAC. Plus exactement, à un instant donné, seule une des deux machines possède ces deux adresses. L'adresse MAC est de la forme : 00:00:5E:00:01:<VRID>. Le VRID, configuré manuellement, est commun aux deux machines : c'est le numéro de l'instance VRRP. Une machine peut participer à plusieurs instances VRRP, par exemple pour secourir deux routeurs différents. Bien entendu, chaque machine dispose également d'une adresse IP unique sur le réseau local.

À un instant donné, seule une machine répond à l'adresse IP virtuelle et à l'adresse MAC correspondant au VRID ci-dessus (c'est le maître), les autres restant en sommeil (ce sont les esclaves). La machine active est celle ayant l'identifiant le plus élevé, l'identifiant étant un entier compris entre 0 et 255 configuré manuellement sur chaque machine participant à l'instance VRRP. La machine active émet toutes les secondes un datagramme multicast (adresse 224.0.0.18) pour indiquer aux autres qu'elle est toujours vivante :

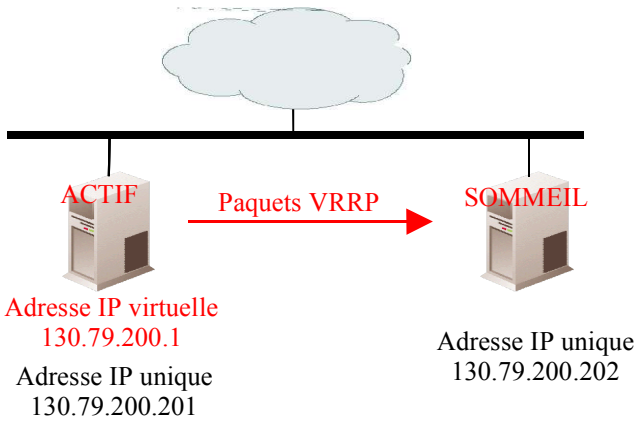


Figure 5 – *VRRP état « optimal »*

Lorsque la machine active défaille, les paquets multicast ne sont plus émis :

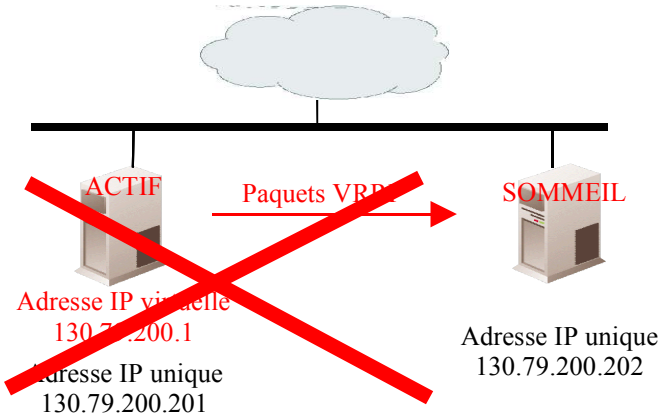


Figure 6 – *Défaillance du serveur actif*

La seconde machine ayant l'identifiant le plus haut détecte l'absence de paquet VRRP : elle s'approprie alors l'adresse IP virtuelle et l'adresse MAC décrite ci-dessus, puis elle émet un paquet « gratuitous ARP» pour chaque adresse IP de l'interface participant au mécanisme VRRP. Ce paquet provoque une mise à jour forcée du cache ARP des autres machines du réseau local, ce qui leur permet de continuer à dialoguer avec l'adresse 130.79.200.202. Ce paquet permet également aux commutateurs de modifier leur table de « forwarding ». La nouvelle machine active émet ensuite, à son tour, les paquets VRRP multicast :

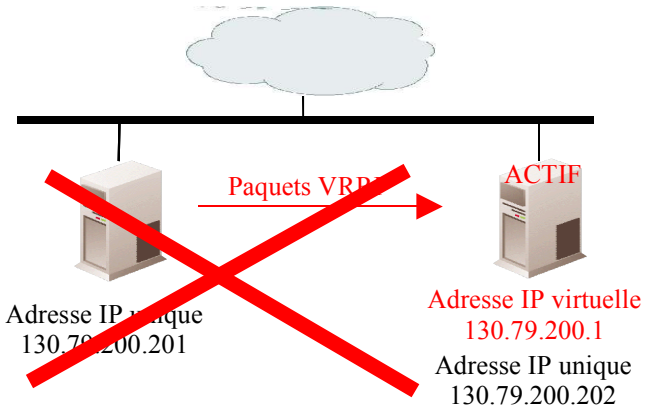


*Figure 7 – Reprise par le serveur en sommeil*

La spécification de VRRP stipule que la détection de défaillance se fait au bout de trois paquets non reçus, ce qui fait que la reprise se passe en moins de quatre secondes.

# 5 Mise en œuvre

## 5.1 Plate-forme

Le serveur DNS principal hébergé au CRC (130.79.200.1) avait été déplacé en décembre 2001 sur une nouvelle machine (bi-Pentium III à 1 GHz, sous FreeBSD [2]). Cette machine héberge, outre le serveur DNS, le relais de messagerie, ainsi que l'antivirus associé pour l'ensemble du réseau Osiris. Correctement dimensionnée pour les services qu'elle offre, elle a donc été intégrée à la nouvelle architecture.

Une seconde machine similaire a été ajoutée pour la redondance. Notre système comporte ainsi deux machines participant à l'instance VRRP. Ces deux serveurs sont bien évidemment localisés dans des lieux géographiques distincts pour minimiser les risques de pannes environnementales.

## 5.2 Implémentation utilisée

L'implémentation de VRRP utilisée est freevrrpd [3], une implémentation libre de VRRP présente dans les « ports » [4] de FreeBSD, qui s'avère robuste et assez complète. La possibilité d'associer des scripts aux changements d'état (par exemple lorsqu'une machine en sommeil devient active) est indispensable dans notre contexte. Il manque toutefois quelques fonctionnalités, décrites dans la RFC 2338, comme l'authentification et le chiffrement : seul le mécanisme « simple text password » est présent.

## 5.3 Comportement des services

Si le principe de VRRP est simple, « the devil is in the detail » : la mise en œuvre concrète recèle de nombreux pièges. Les services présents sur ces machines doivent se comporter de manière différente en fonction de l'état VRRP (actif ou en sommeil).

Les deux services associés à l'adresse IP publique 130.79.200.1 sont :
– le service de résolution de noms DNS ;
– le relayage de messagerie et l'antivirus associé.

Le problème commun à ces deux services est la prise en compte de la nouvelle adresse IP virtuelle lorsque l'esclave devient maître. De plus, lorsque la machine de secours repasse dans l'état « esclave », le service de relayage de messagerie doit se comporter de manière différente pour transmettre tous les messages en instance au relais officiel. La suite décrit en détail ces problèmes et les solutions apportées.

Le service DNS lié à l'adresse 130.79.200.1 est le démon « bind » [5]. Lorsqu'il se lance, il s'attache par défaut à toutes les interfaces réseau présentes sur la machine, ce qui pose problème lorsqu'une machine en sommeil devient active car la nouvelle adresse IP virtuelle n'est pas prise en compte. La solution, simple, consiste à ne démarrer le démon que lorsque la machine devient active.

Le relayage de messagerie utilise, quant à lui, le démon « sendmail » [6]. Le problème est plus complexe que précédemment : lorsque la machine en sommeil devient active, elle reçoit des messages, dont certains aboutissent dans la file d'attente. Lorsqu'elle revient en sommeil, les messages vont partir avec une adresse IP source différente de celle du relais officiel, ce qui peut poser des problèmes vis-à-vis des règles de filtrage en vigueur par exemple dans des laboratoires. Pour rendre le mécanisme transparent pour tous, la solution adoptée est d'avoir deux comportements différents (donc deux `sendmail.cf` différents [7]) sur la machine de secours :
– lorsqu'elle est active, elle se comporte comme le relais normal qu'elle remplace ;
– lorsqu'elle revient en sommeil, elle vide sa file d'attente en envoyant tous les messages vers le relais normal.

Le problème d'attachement à l'adresse officielle évoqué ci-dessus pour « bind » se pose de la même manière, mais nécessite en revanche une modification de la configuration de « sendmail ».

La figure 8 illustre le comportement du système lorsque tous les éléments sont opérationnels :

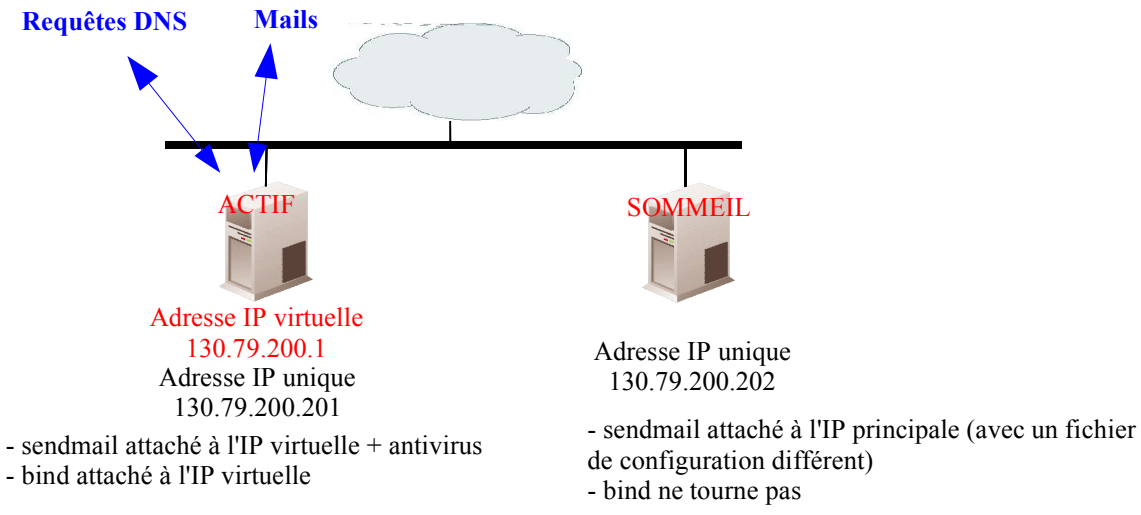


Figure 8 – *Comportement des services en état optimal*

La figure 9 illustre le comportement du système lorsque le serveur principal ne répond plus :

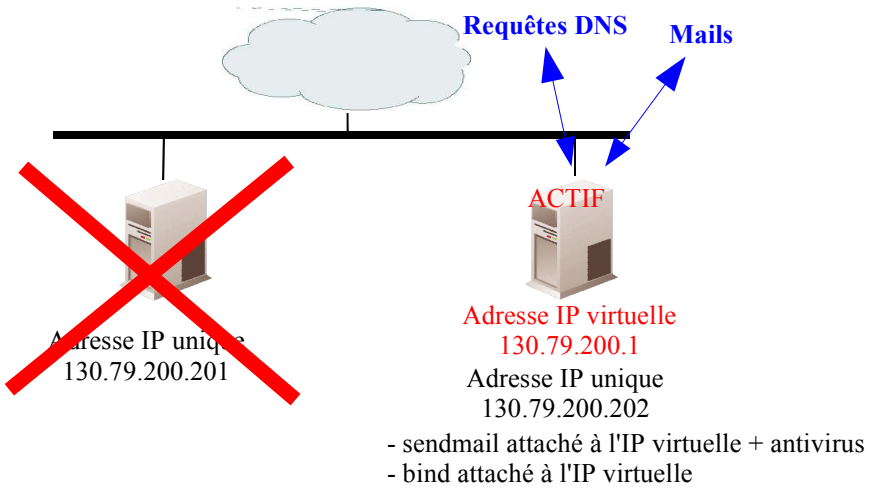


*Figure 9 – Comportement des services en mode dégradé*

La figure 10 illustre le comportement du système lors du retour du serveur principal dans l'état actif :

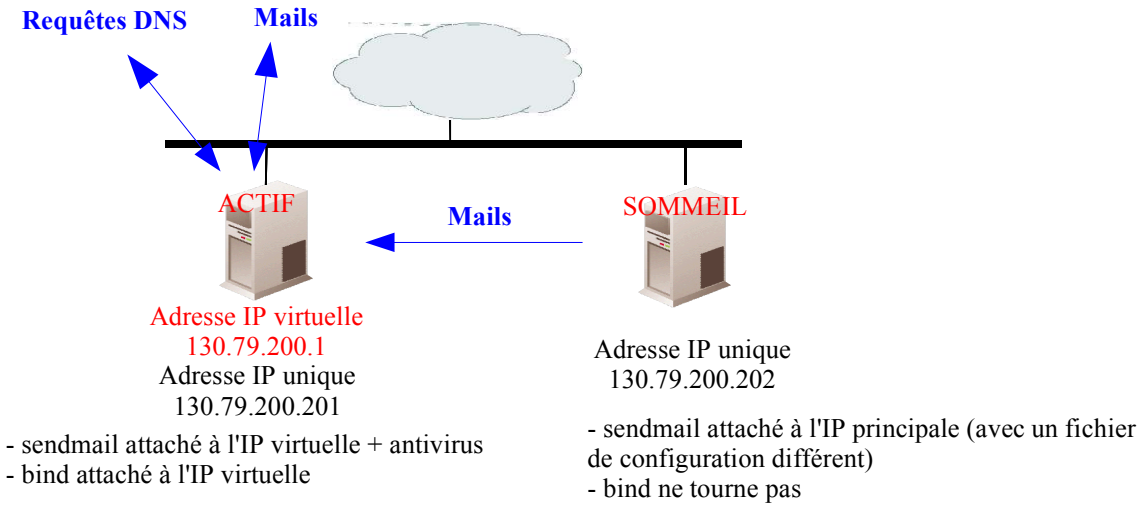


Figure 10 – *Comportement des services lors de la reprise du serveur principal*

## 5.4 Scripts associés

Grâce à l'implémentation « freevrrpd », des scripts ont été associés aux changements d'état VRRP pour réaliser les modifications de comportement évoquées précédemment.

Sur le serveur principal, un seul script est utilisé pour le passage à l'état « actif », pour lancer « bind » et « sendmail » attachés à l'adresse IP virtuelle. Il n'y a pas besoin de script pour passer en état « sommeil », puisque cette machine est soit active, soit en panne.

Sur le serveur secondaire :
- le script de passage à l'état « actif » est comparable à celui du serveur principal ;
- le script de passage à l'état « sommeil » arrête « bind » et relance « sendmail » avec un fichier de configuration différent pour vider la file d'attente.

Enfin, la synchronisation des informations (`sendmail.cf`, `named.conf`, zones DNS, antivirus et signatures associées) entre le serveur principal et le serveur secondaire est très importante, aussi un script a été rédigé pour transférer les informations, en cas de modification détectée par un changement de signature MD5 d'une partie de l'arborescence.

### 5.5 Mise en exploitation opérationnelle

La mise en place s'est déroulée en 2 étapes :
- 30 avril 2003 : l'adresse du serveur DNS principal a été modifiée et l'adresse « officielle » est devenue l'adresse IP virtuelle ; le comportement de « sendmail » et de « bind » avec cette adresse virtuelle a été validé en production ;
- 2 mai 2003 : le deuxième serveur a été mis en service, « freevrrpd » a été mis en place sur les deux machines et l'ensemble a été mis en production.

Les deux étapes ont provoqué une interruption de service (cumulée) de moins de 6 minutes. La mise en place du nouveau serveur redondant s'est donc faite de manière presque transparente.

L'annonce de la suppression programmée des serveurs de noms délocalisés sur les campus a été faite aux correspondants réseau lors de la réunion du 24 juin 2003. Cette suppression sera effective au 30 septembre 2004.

## 6 Conclusion

La conclusion de cet article pourrait se résumer à une valeur : 99,99997 % de disponibilité du service DNS sur les 5 mois écoulés depuis la mise en production. Les 0,00003 % d'indisponibilité sont dûs à une mise à jour du système d'exploitation.

Avec une telle valeur, il est difficile de trouver des inconvénients à l'architecture mise en place. Toutefois, puisqu'il est d'usage d'en trouver, il faut mentionner que le réseau sous-jacent doit être en adéquation avec l'objectif de disponibilité, et donc d'une fiabilité sans faille. Par ailleurs, le dispositif présenté est basé sur une détection de défaillance au niveau de la couche IP et non de la couche applicative DNS : cela reste à améliorer, et des pistes sont d'ores et déjà envisagées.

Sur le plan des avantages, la disponibilité est bien sûr évidente. Avec la suppression en 2004 des anciens serveurs délocalisés, la charge en terme d'administration système diminuera significativement. Enfin, la conséquence de ce dispositif est la facilité de maintenance matérielle ou logicielle sans interrompre le service, et donc sans avoir la pression associée lors d'interventions aussi critiques.

Enfin, en tant qu'administrateur de ces serveurs, je peux dire que j'aborde les mises à jour de manière beaucoup plus sereine... ;-)

## Références


[1] RFC 2338 : Virtual Router Redundancy Protocol, avril 1998
[2] FreeBSD : http://www.freebsd.org
[3] freevrrpd : http://www.bsdshell.net
[4] FreeBSD Handbook : http://www.freebsd.org/doc/en_US.ISO8859-1/books/handbook/ports.html
[5] ISC BIND : http://www.isc.org/products/BIND
[6] Sendmail : http://www.sendmail.org
[7] Pierre David, Jacky Thibault et Sébastien Vautherot. Kit Jussieu : http://www-crc.u-strasbg.fr/docs/kit-jussieu