

# Un réseau de campus résilient à moindre coût (ou les anneaux au secours des étoiles)

## Pascal Mouret

DOSI Campus Luminy / Université d'Aix-Marseille  
163 avenue de Luminy  
13288 Marseille cedex 9

## Résumé

*Au fil des ans, le réseau devient une ressource de plus en plus critique, ne serait-ce que du fait de la généralisation de la virtualisation, de la centralisation d'applications, ou de la ToIP. Le besoin accru de disponibilité qui en découle a bien évidemment des implications importantes sur la gestion des réseaux, par rapport, entre autres, aux dysfonctionnements éventuels ou aux évolutions programmées.*

*Pour adresser cette problématique, nous avons travaillé depuis plusieurs années sur la mise en œuvre de résilience, avec une contrainte de « moindre coût ». Dans cette optique, les architectures en anneau sont apparues comme étant une solution très intéressante. En effet, il suffit d'un seul lien supplémentaire pour transformer une topologie existante en étoile, quel qu'en soit le nombre de branches, en une topologie en anneau et pour ainsi augmenter significativement la disponibilité du réseau.*

*A la lumière de notre expérience, nous exposons ici différentes solutions disponibles ainsi que leurs corollaires.*

## Mots-clefs

*réseau de campus, résilience, moindre coût, étoile, anneau*

## 1 Introduction

Au fil du temps, le réseau devient une ressource de plus en plus critique. Plusieurs éléments concourent à cet état de fait : les nouvelles applications telles que la téléphonie sur IP ou les TICE (visioconférence ou cours à distance par exemple), la (re)centralisation des ressources applicatives et de stockage ou de plus en plus la virtualisation complète du poste de travail. Tous ces éléments nécessitent une disponibilité sans (trop de) failles, et les attentes des utilisateurs sont de plus en plus importantes. Deux aspects sont à considérer dans cette « haute disponibilité » : la panne bien évidemment, mais aussi et surtout la maintenance, qu'on a parfois tendance à sous-estimer dans des réflexions de ce type. Or, faire fonctionner un réseau implique de le mettre à jour, de le reconfigurer, de le remodeler et il est important de pouvoir le faire de manière transparente pour les utilisateurs. Trop souvent, la nécessité de maintenir en fonctionnement le réseau et les risques de perturbations que représentent ces actions préventives peuvent inciter à les négliger, exposant de ce fait l'infrastructure à d'autres types de problèmes.

Depuis une petite décennie, les réseaux opérateurs, puis les réseaux métropolitains ont commencé à s'adapter [1], mais la pression commence maintenant à s'accroître sur les extrémités de ces réseaux et les réseaux de campus notamment.

Forts de ce constat, nous avons travaillé depuis plusieurs années sur les techniques de continuité de service. Compte tenu des budgets parfois tendus, l'idée a été dès le départ de voir comment on pouvait, « à moindre coût », adapter une architecture réseau existante pour en augmenter la disponibilité, afin d'avoir un réseau le plus stable possible, facilement maintenable, et supportant des événements, planifiés ou non, sans impact (trop) visible par les utilisateurs. Depuis deux ans, avec la prise de conscience de l'intérêt que pouvaient avoir les protocoles de gestion d'anneaux, ces idées ont pu être progressivement mises en œuvre, et cet article se fonde sur l'expérience acquise à cette occasion. Après une introduction sur la résilience et l'importance de l'apport de telles architectures, nous présentons différentes techniques disponibles pour atteindre ce but. Puis, nous décrivons l'architecture réseau mise en place sur notre campus avec les problèmes rencontrés, les solutions apportées, et les enseignements que nous en avons retirés. Enfin, nous nous

intéressons à deux corollaires importants que cette expérience nous a fait découvrir : peut-on avoir une architecture résiliente sans supervision ? et quand bien même chacun des protocoles utilisés est simple, peut-on réellement gérer un réseau de plusieurs dizaines d'équipements avec des interactions importantes entre eux sans outil adapté ?

## 2 La résilience

« La résilience est la capacité d'un système ou d'une architecture réseau à continuer de fonctionner même en cas de panne ».

Aujourd'hui encore plus qu'hier, un utilisateur qui n'a plus de réseau ne peut plus travailler. La plupart des applications métier sont maintenant centralisées avec des accès distants pour les utilisateurs. Cela apporte d'énormes avantages, entre autres en termes de facilité d'administration, de regroupement et de cohérence des données, mais en cas de coupure réseau, il n'est plus possible aux utilisateurs d'y accéder. De même, ils ne peuvent souvent même plus accéder à leurs propres données, stockées sur un serveur sécurisé distant, ni même téléphoner.

Une coupure ou une perturbation réseau peuvent avoir des conséquences variables selon leur ampleur ou leur durée. Autour de la minute, il n'y a guère que la téléphonie sur IP qui sera impactée, avec des phénomènes auditifs voire une coupure des connexions pour les durées plus longues. Au bout de quelques minutes seulement, il peut déjà y avoir des pertes de données du fait de l'expiration des délais d'attente de certains protocoles. Quelques dizaines de minutes suffisent à provoquer des coups de fil aux équipes d'assistance (pour ceux qui ont un téléphone autre qu'un téléphone IP !) ainsi que des premiers signes d'impatience, notamment s'il y a des cours, ou si on est en période d'inscription, etc. Au delà d'une demi-heure une heure, on peut s'attendre à voir les utilisateurs s'en aller.

D'autre part, la plage horaire d'utilisation du réseau s'allonge. La multiplicité d'applications nécessitant le réseau fait qu'à toute heure du jour ou de la nuit, il y a des personnes qui souhaitent accéder à une application ou à des données. Les créneaux de maintenance sont de plus en plus difficiles à trouver, sauf à être insomniaque (encore que ...).

Cependant, les interruptions de fonctionnement sont inévitables. Sur 120 équipements, si l'on suppose que l'on en renouvelle 20 % chaque année, cela correspond à 24 changements par an. Il y a également des mises à jour de sécurité à faire, qui nécessitent souvent un redémarrage. Pour éviter les interruptions, on a souvent tendance à repousser les mises à jour qui en découlent, mais au final, les équipements sont de plus en plus susceptibles de tomber en panne ou de subir une attaque, ce qui ne va pas dans le sens d'une diminution de la durée d'interruption. Et quelles que soient les mesures prises pour les éviter, il y aura aussi des pannes.

Mettre en place de la résilience, c'est permettre aux utilisateurs de pouvoir disposer normalement de leur outil de travail indépendamment de l'endroit où sont leurs données ou leurs applications et au moment où ils en ont besoin. Côté exploitation, cela veut dire sortir du mode « pompier » en cas de panne, diminuer grandement le stress sur les interventions (planifiées ou non), et globalement fluidifier la gestion du réseau.

De façon générale, la résilience permet de décorrélérer l'utilisation du réseau qu'en font les utilisateurs des actions de maintenance préventive et curative inhérentes au fonctionnement de ce réseau.

## 3 De l'étoile à l'anneau, il n'y a qu'un lien

Dans une architecture en étoile, le temps d'indisponibilité, du point de vue d'un utilisateur, est fonction des temps d'indisponibilité des éléments qui le séparent des ressources auxquelles il a besoin d'accéder. Or, le seul équipement dont ils dépendent obligatoirement, c'est celui sur lequel ils sont connectés. Fonctionnellement, c'est le seul avec lequel leur équipement terminal « dialogue ». Tout le reste de la transmission des données se fait de manière complètement transparente pour ce dernier. Par conséquent, en mettant en place une architecture adaptée, on peut faire en sorte que le chemin soit recalculé en cas de perte d'un élément, rendant ainsi l'utilisateur moins sensible aux événements survenant sur le réseau. Dans ce cas-là, la durée d'indisponibilité du réseau pour un utilisateur devient la durée d'indisponibilité de son équipement de raccordement augmentée, dans le pire des cas, du cumul des temps de calcul d'un nouveau chemin en cas d'incident. L'objectif est que ce dernier cumul soit bien inférieur aux temps d'indisponibilité des équipements. Cela nécessite des architectures et protocoles adaptés, afin de minimiser le temps de calcul sans augmenter la complexité jusqu'à un point où elle peut elle-même être source de problèmes.

On passera rapidement sur la solution qui consiste à doubler tous les équipements et à relier les paires ainsi constituées

entre elles par des doubles attachements. Cela répond tout à fait à l'objectif de disponibilité. Également, avec des protocoles appropriés (on en parlera plus loin), on peut gérer chaque paire d'équipements comme un seul équipement plus gros et conserver ainsi l'architecture en étoile classique. Mais le coût mis en jeu est énorme aussi bien en investissement qu'en renouvellement.

Si on veut minimiser les coûts, on peut se poser la question de savoir s'il n'y a pas une autre architecture qui permet d'atteindre cet objectif d'avoir deux chemins distincts vers la ressource considérée. Et la conclusion est que l'anneau est l'architecture optimale correspondant à cette contrainte. En effet, il suffit d'un seul lien supplémentaire pour transformer une topologie en étoile existante, quel qu'en soit le nombre de branches ou de niveaux, en une topologie en anneau.

Dans un anneau, tous les nœuds ont exactement deux chemins distincts vers chacun des autres nœuds, donc on est bien en mesure d'atteindre les deux objectifs fixés : la résilience et le surcoût minimum.

Cela dit, rajouter un lien n'est pas toujours (aisément) faisable :

- quel que soit le nombre  $n$  de nœuds, on n'utilise au total que deux ports de plus par rapport à l'architecture en étoile, mais leur répartition est différente ; chaque équipement doit disposer de deux ports pour l'interconnexion au lieu d'un pour les équipements feuilles et  $n-1$  pour l'équipement central d'une étoile ;
- les liaisons vont être montées différemment ; il faut notamment pouvoir disposer de liaisons transversales directes (entre les équipements feuilles) ou au pire indirectes (en utilisant les liens vers le cœur et en jarretiérant à ce niveau). Cela peut augmenter les distances, et nécessiter d'autres technologies plus onéreuses (1000baseLX au lieu de 1000baseSX ou 10GbaseLR au lieu de 10GbaseSR par exemple).

Au niveau performance, il n'y a aucun impact jusqu'à 3 nœuds. Par contre, au delà, les débits maximum vont être fonction de la répartition du trafic et des liens. Également, la boucle n'est pas une architecture naturelle dans les réseaux Ethernet, et son utilisation nécessite des outils supplémentaires que nous allons voir maintenant.

## 4 Protocoles de gestion de boucles

### 4.1 Niveau 2

#### 4.1.1 Cas général

La gestion des boucles de réseau au niveau 2 est une problématique bien connue, pour laquelle on dispose d'une solution depuis longtemps. Il s'agit du STP (Spanning Tree Protocol), dont l'algorithme remonte à 1985 et qui a été normalisé en 1990 par l'IEEE (802.1D). Il a connu de nombreuses évolutions depuis, propriétaires ou normalisées, comme le RSTP (Rapid STP, 802.1w-1998, inclus ultérieurement dans 802.1D-2004), ou le MSTP (Multiple STP, 802.1s, 2002), mais, fondamentalement, son fonctionnement reste le même.

Nécessitant jusqu'à 50s pour calculer une nouvelle topologie et reprendre la transmission des trames de données, l'algorithme initial a été grandement amélioré et le MSTP permet aujourd'hui des temps de convergence de l'ordre de quelques secondes dans le pire des cas. Toutefois, si le spanning-tree reste actuellement sans égal, et indispensable à ce titre, pour la détection des boucles involontaires, il est nettement moins adapté pour la gestion des boucles volontaires.

En effet, l'arbre qu'il construit est global et toute modification du réseau est susceptible d'entraîner un recalcul de l'arbre, même si elle n'intervient pas sur les liens d'« infrastructure ». Le temps de convergence est fonction de la taille du réseau. Gérer un spanning-tree sur une « grosse » infrastructure nécessite une ingénierie très lourde et beaucoup de configuration à faire sur chaque équipement, la moindre erreur pouvant entraîner des instabilités sur l'ensemble du domaine couvert[3].

Conscient de ces différentes problématiques, l'IEEE a récemment normalisé le SPB (Shortest Path Bridging, 802.1aq, 2012), dont un des objectifs est de remplacer le spanning-tree et de proposer des architectures sans lien bloqué. Aujourd'hui, il ne semble pas y avoir beaucoup d'implémentations mais cela reste certainement à surveiller de près.

En attendant, il existe une autre alternative très intéressante. Depuis plusieurs années, la problématique des anneaux s'est posée notamment au niveau des réseaux opérateurs puis des réseaux métropolitains. Les gros réseaux de collecte utilisaient traditionnellement des structures en anneau, basées sur SONET par exemple, qui permettaient une convergence dans des temps de l'ordre de 50ms en cas de problème. Avec la popularisation d'Ethernet et sa

généralisation dans ces réseaux, il était nécessaire de disposer d'une solution permettant un niveau de service comparable à ce qui existait auparavant. De nombreux efforts ont été faits au travers de protocoles propriétaires ou de normalisations pour atteindre ce résultat. Quasiment tous les acteurs majeurs du réseau proposent leur solution, pour certains réservée aux équipements pour réseaux métropolitains ou réseaux opérateurs, pour d'autres disponible également jusque dans les équipements de distribution.

Il est difficile de les citer toutes, mais pour s'en tenir aux standards, il y en a pour le moment trois.

RPR (Resilient Packet Ring, 802.17) normalisé par l'IEEE dès 2004. Sa dernière révision date de 2011 (802.17d). L'approche retenue par l'IEEE a consisté à développer une nouvelle couche de liaison tirant parti des architectures en anneau sans désactiver de lien et permettant une convergence de l'ordre de 50ms. Il remplacerait Ethernet ou plutôt serait une sous-couche à Ethernet. Il nécessite des équipements spécifiques et est clairement orienté opérateur.

ERPS (Ethernet Ring Protection Switching, recommandation G.8032 de l'ITU-T, 2008) parfois nommé aussi R-APS (Ring Automatic Protection System). Une deuxième version G.8032 v2 a été approuvée en 2009. Il s'agit d'un protocole au dessus d'Ethernet qui est implémentable sans changement majeur des équipements. Cisco et Juniper notamment l'ont intégré dans une partie de leur gamme.

Enfin, EAPS (Ethernet Automatic Protection System), faisant l'objet du RFC3619 (statut « informational ») de 2003. Il a été initié par Extreme Networks, mais d'autres constructeurs proposent leur protocole s'appuyant sur ce RFC : HP (au travers de 3Com/H3C) avec RRPP ou Allied Telesis avec EPSR (à ne pas confondre avec ERPS !).

ERPS et EAPS ont des fonctionnements similaires, et c'est une implémentation de ce dernier que nous avons mise en œuvre : RRPP (Rapid Ring Protection Protocol). L'un comme l'autre permettent une convergence de l'ordre de 50 à 250ms, ce qui permet de passer complètement inaperçu au niveau de la ToIP. Nous utiliserons les dénominations de RRPP pour détailler la suite, mais le principe est le même.

Sans trop rentrer dans les détails du protocole, le point clef consiste à avoir constamment le long de l'anneau un lien sur lequel le trafic de données est bloqué, afin d'éviter les boucles. Un commutateur particulier de l'anneau, le « master », a cette charge qu'il accomplit en activant ou désactivant le trafic de données sur un de ses deux ports situé sur l'anneau, nommé « port secondaire ». Le mécanisme de détection est basé sur une notification des autres commutateurs de l'anneau ou sur des paquets « hello » envoyés sur un VLAN particulier, nommé « control VLAN ». A chaque changement d'état de son lien, il envoie un message aux autres nœuds pour qu'ils rafraîchissent leurs tables de commutation (MAC et ARP/ND).

RRPP a le double avantage d'être relativement simple, avec une configuration qui se fait aisément, et de pouvoir fonctionner avec des équipements non-RRPP. Ces derniers doivent seulement transmettre le VLAN de contrôle et les VLANs utilisateurs sur leurs deux ports de l'anneau. En cas de changement d'état de l'anneau, le fonctionnement reste le même pour les équipements RRPP. Les équipements non RRPP en revanche ne peuvent pas bénéficier des informations transmises. De fait, en cas de changement de topologie, leur convergence sera fonction de la vitesse à laquelle leurs tables de commutation sont rafraîchies.

Il permet une grande variété de configuration avec des anneaux qui peuvent se recouvrir totalement ou partiellement, aussi bien pour faire du partage de charge que pour étendre l'architecture au delà du nombre de nœuds maximum d'un anneau[4]. Également, point important, RRPP isole les domaines de spanning-tree. Le spanning-tree doit être désactivé sur les liens RRPP et la résilience y est assurée exclusivement par RPPP. Le spanning-tree peut être conservé pour la détection de boucles accidentelles locales aux nœuds.

#### **4.1.2 Cas de 2 nœuds seulement**

Lorsqu'il y a deux nœuds seulement, on peut gérer l'ajout du lien comme vu précédemment, mais on dispose également de solutions beaucoup plus simples ou performantes.

Tous les matériels proposent une notion de « lien de secours ». Il s'agit simplement de garder un lien non actif et de le réactiver dès que l'on détecte que le lien principal est tombé. C'est le cas le plus simple d'un protocole d'anneau et cela ne nécessite de configuration que sur un seul équipement. Les deux liens peuvent avoir des débits différents.

Mais il y a aussi et surtout l'agrégation de lien. Dans ce cas-là, les deux liens sont actifs et le trafic est réparti sur ces deux liens, ce qui permet d'augmenter la bande passante, et de maintenir le fonctionnement, éventuellement dégradé, en cas de perte d'un lien. Elle peut être statique ou dynamique. La différence majeure entre ces deux modes est que le choix des ports actifs dans une agrégation se fait selon des critères exclusivement locaux pour l'agrégation statique, alors qu'il

est le résultat d'une négociation pour l'agrégation dynamique. Cette dernière met en jeu le protocole LACP (Link Aggregation Control Protocol, IEEE 802.3ad, 2000, intégré plus tard dans 802.1AX-2008). Malgré les problèmes inhérents à une communication protocolaire, elle est beaucoup plus robuste que l'agrégation statique, le trafic n'étant transmis que lorsqu'il y a au moins deux ports, un de chaque côté d'un lien, dont les configurations sont cohérentes.

Dans tous les cas, on ne peut avoir d'actifs, à un moment donné, que des ports de même débit et même duplex. Par contre, LACP permet de définir au sein d'une agrégation des ports ayant des débits ou modes différents. A tout moment, les ports actifs simultanément dans l'agrégation auront les mêmes débits et modes, mais en cas de problème sur ces ports, un groupe de ports de débits ou modes différents pourra être activé.

Attention toutefois : le débit maximum d'une agrégation de lien dépend énormément du mode de répartition du trafic choisie (celui-ci est basé sur une clef de hashage en fonction de l'adresse IP source ou destination ou de l'adresse Mac source ou destination entre autres). En fonction de l'utilisation du lien, il conviendra de vérifier que le mode choisi ne privilégie pas un lien plutôt que les autres.

## 4.2 Niveau 3

La boucle est moins problématique au niveau 3 qu'au niveau 2. Il n'y a en fait pas de notion de gestion d'anneaux à proprement parler. La problématique principale se situe au niveau de l'accès à la passerelle par défaut.

Une solution traditionnelle à ce problème est le protocole VRRP (Virtual Router Redundancy Protocol). Défini initialement dans le RFC 2338 (1998), il en est actuellement à sa version 3 (RFC 5798, 2010). Le principe consiste à partager une adresse IP « virtuelle » entre deux équipements ou plus. A tout moment, un seul de ces équipements utilise cette adresse. C'est elle qui sera la passerelle par défaut des équipements terminaux. Cette approche marche bien mais la configuration doit être faite pour tous les VLANs, ce qui peut être assez fastidieux. Également, si on fait du filtrage terminal, il faut le dupliquer sur les différents boîtiers concernés. Enfin, si tous les VLANs ne sont pas configurés sur tous les équipements, alors du routage dynamique est nécessaire et complexifie encore plus l'architecture.

Une autre approche paraît beaucoup plus intéressante : le « stacking ». Le stacking en lui-même n'est pas nouveau, mais ce qui est beaucoup plus récent, c'est la possibilité de le faire au travers de liens Ethernet.

Le stacking recouvre un grand nombre de significations, depuis l'administration par une adresse IP unique d'équipements multiples jusqu'à la création de commutateurs-routeurs virtuels. C'est cette dernière notion qui nous intéresse. En effet, d'un point de vue logique, cela revient à configurer en une action unique de multiples boîtiers. Lorsque le stacking est faisable au travers des liens Ethernet, cela veut dire que l'on peut monter un commutateur virtuel à cheval sur un territoire géographique assez vaste. Par principe, le stacking ne nécessite pas un anneau mais cette architecture est recommandée. Outre la configuration simplifiée, l'avantage du commutateur virtuel est qu'en cas de panne d'un des boîtiers, toutes les fonctions qu'il assurait sont reprises par les membres restants. La bascule se fait généralement en quelques dizaines de millisecondes. Également, il permet de faire de l'agrégation de lien distribuée et donc de ne plus dépendre d'un seul boîtier.

Le stacking est toujours propriétaire et n'est souvent possible qu'avec des équipements homogènes (du fait qu'il dépend beaucoup de l'architecture interne des équipements et des possibilités de leurs OS respectifs). Heureusement, le cœur de réseau, où sont généralement concentrés les équipements de niveau 3, est souvent le point du réseau le plus homogène.

## 5 Mise en œuvre

### 5.1 L'historique

Notre campus est composé d'une dizaine de bâtiments, répartis entre 4 UFR (plus une dizaine d'autres bâtiments gérés par des laboratoires ou organismes de recherches). A l'origine, le réseau était architecturé en étoile autour de 6 équipements de niveau 3 reliés en 10Gb/s : un en cœur et cinq en périphérie (un par UFR plus un pour les laboratoires et autres bâtiments hors UFR). Cette architecture se déclinait au niveau 2 également en étoile au sein de chaque UFR (ou groupe de bâtiments), pour un total d'environ 120 équipements.

Cette architecture est relativement simple et maîtrisée. Cependant, tous les équipements et liens de concentration sont critiques. En vertu des réflexions exposées plus haut, il fallait corriger cela, et l'idée était de le faire à moindre coût, en

réutilisant au maximum les équipements déjà en place.

L'idée directrice était que l'arrêt d'un équipement ne devait avoir d'impact que pour les équipements terminaux qui étaient connectés directement dessus. Autant que possible, tous les serveurs devaient être reliés de façon transparente à deux équipements différents.

## 5.2 Redondance du cœur

Après étude, la première étape a consisté à agréger les 6 commutateurs-routeurs en un seul virtuel. Cela a été réalisé fin 2011 début 2012. Nous avons mis en place le protocole IRF (Intelligent Resilient Framework) disponible sur nos équipements de cœur (tous de la même gamme HP). Le stacking s'est fait au travers des liens 10Gb/s déjà présents (plus un supplémentaire pour fermer l'anneau). Toute la commutation ainsi que toutes les opérations gérées par les ASIC se font en local sur chacun des boîtiers. Toutes les fonctions avancées (non gérées par les ASIC) sont effectuées par un seul boîtier, le master. Toutes les modifications de configuration sont répercutées sur l'ensemble des boîtiers. En cas de perte du master, toutes les fonctions avancées sont reprises par le boîtier disponible le plus prioritaire. En cas de perte de lien ou du master, la reprise du fonctionnement se fait en moins de 50ms.

Au départ, l'idée était de faire un commutateur virtuel constitué par les 6 commutateurs-routeurs. Cependant, la configuration étant appliquée exactement de la même manière sur tous les boîtiers, et n'importe quel boîtier étant susceptible de réaliser tout le traitement logiciel à la place du master, cela voulait dire que l'on s'alignait sur les capacités du boîtier le plus faible. En l'occurrence, nous avons des ACLs sur tous les boîtiers et en les cumulant (tous les boîtiers stockaient maintenant toutes les ACLs, puisqu'ils avaient tous exactement la même configuration), nous dépassions la capacité du boîtier le plus faible (paradoxalement, le boîtier d'agrégation 10Gb/s). Nous avons alors décidé de scinder notre IRF en deux IRF reliés entre eux par un lien LACP. Pour 6 commutateurs (et donc 5 liens), il nous a fallu rajouter 3 liens au total. Fonctionnellement, hormis le fait de gérer deux configurations distinctes, on atteint sensiblement le même objectif.

Les serveurs ont été rattachés par des agrégations de liens (LACP) partagées sur deux boîtiers d'un même IRF.

En termes de performances, pas d'impact notable : un test à base de ping réalisé avant et après la migration a montré une très légère augmentation du temps (de l'ordre de 0,1ms) pas forcément significative. Au niveau CPU, on constate également une activité légèrement plus importante sur le master par rapport aux slaves.

En termes de stabilité, le bilan est largement positif, même s'il nous est arrivé d'avoir un problème majeur du type « split-brain » où l'un des boîtiers perd les liens IRF avec ses voisins, mais continue de fonctionner. Dans ce cas, on se retrouve avec deux instances distinctes du même commutateur virtuel avec les conséquences que cela peut avoir sur les agrégations de liens créées à cheval sur les deux ensembles.

Enfin, en termes d'administration, la gestion de tous les jours est réellement facilitée. On ne gère plus que 2 routeurs au lieu de 6. Par contre, la migration a été relativement complexe parce que nous étions sur un réseau fonctionnel que nous ne souhaitions pas interrompre. En effet, la création d'un IRF implique plusieurs redémarrages : pour la mise à jour des boîtiers (tous doivent avoir la même version de système), la renumérotation des boîtiers (chaque équipement dispose d'un numéro, 1 par défaut, qui doit être unique au sein d'un IRF) et l'adjonction d'un boîtier à un IRF existant (il peut être réduit à un master). Également, la dénomination des ports contient une référence au numéro de boîtier et aucune des opérations précédentes n'adapte la configuration. Par conséquent, si l'on n'a pas pris les dispositions adéquates, il faut s'attendre à avoir un boîtier non fonctionnel après redémarrage. Autant dire qu'une préparation minutieuse est primordiale. Autant que possible, il vaut mieux prévoir l'IRF (ou équivalent) avant la mise en place des équipements.

Un dernier point à considérer est le problème des futures mises à jour. Tous les boîtiers doivent avoir exactement la même version du firmware. Redémarrer en une seule fois l'ensemble du commutateur virtuel n'apporte pas grand chose par rapport à auparavant. Un certain nombre de constructeurs proposent un mécanisme d'ISSU (In-Service Software Upgrade). Cette notion ne regroupe pas toujours tout à fait les mêmes fonctionnalités selon les cas. Ici, il s'agit de faire des redémarrages décalés. Un premier commutateur de l'IRF, autre que le master, redémarre sur la nouvelle version et forme un IRF nouvelle version dont il est le master. Puis les commutateurs de l'ancien IRF redémarrent les uns après les autres de sorte qu'il n'y en ait jamais qu'un seul à la fois qui soit en cours de redémarrage. Cela permet aux agrégations de lien de toujours être fonctionnelles. Le mécanisme d'ISSU ne peut être utilisé que lorsque la version de firmware courante et la nouvelle sont compatibles, ce qui n'était pas notre cas lors de la mise à jour de nos IRF. Au final, c'est une procédure relativement lourde et, a priori, difficilement utilisable. C'est le seul point noir pour le moment.

## 5.3 Le casse-tête de la distribution

Côté distribution, le choix s'est porté sur le protocole RRPP. Les raisons de ce choix ont été, entre autres, une certaine robustesse, la possibilité de fonctionner avec des équipements non RRPP, ainsi que le temps de convergence de l'ordre de 50ms.

Le principe paraissait simple : chaque armoire de distribution correspondait à un anneau distinct, soit de l'ordre de 3 à 4 commutateurs par anneau. Au niveau performances, on optimisait les débits là où il n'y avait qu'un lien vers le cœur, et on restait au même niveau dans le cas où il y en avait déjà deux. En termes de coût, cela impliquait au pire un lien fibre vers le cœur.

Seul problème : pour 3 bâtiments sur 4, il n'y avait qu'un seul commutateur de l'IRF. Donc, les boucles arrivaient forcément sur un seul point de l'IRF qui devenait lui-même un SPOF (Single Point of Failure). Pour le contourner, nous aurions pu prolonger le lien fibre supplémentaire jusqu'à un bâtiment voisin disposant d'un commutateur IRF, mais cela imposait une homogénéité des fibres en interne au bâtiment et en inter-bâtiments, ce qui n'était pas le cas pour nous.

La solution est venue de la possibilité de créer des « sous-anneaux ». Nous avons profité du fait que, dans chaque bâtiment, nos équipements de niveau 3, constituants de l'IRF, étaient dans un local commun à des commutateurs de distribution. Nous avons ainsi pu constituer une boucle principale à cheval sur deux bâtiments en incluant les 2 boîtiers correspondants de l'IRF plus un commutateur de distribution dans chaque local. Cela a nécessité un lien supplémentaire entre les deux commutateurs de distribution. Cette boucle ayant vocation à pallier la perte d'une liaison 10Gb/s, il nous a semblé nécessaire de la passer elle-même entièrement en 10Gb/s. Le surcoût est faible à l'intérieur des locaux techniques mais non négligeable en inter-bâtiments.

A l'inverse de l'IRF, la mise en œuvre de RRPP sur un réseau en production est relativement aisée. En termes d'architecture, on a systématiquement choisi le master à l'opposé du lien le plus important. En effet, en RRPP, c'est le master qui ouvre la boucle en filtrant le trafic de données sur son port secondaire. Ainsi, on optimise le trafic en le faisant transiter via le lien le plus important. La configuration est très simple, et la mise en œuvre l'est tout autant. Le principe consiste à construire physiquement le lien qui referme la boucle, et à le couper administrativement. Ensuite, on rentre la configuration RRPP sur tous les équipements et on l'active. L'anneau va se trouver en état « incomplet ». Il suffit alors de réactiver le lien pour passer l'anneau dans l'état « complet ».

Le problème le plus délicat à gérer est celui de la cohérence des configurations des différents commutateurs constituant l'anneau. En effet, les paramètres du domaine RRPP doivent être strictement les mêmes (VLANs de contrôle, VLANs à protéger et timers notamment) sous peine de créer une boucle et de faire s'effondrer le réseau. De même, les VLANs à protéger autorisés sur les liens doivent être strictement les mêmes. Sinon, dans le meilleur des cas, un VLAN n'est pas secouru lors d'un problème, et dans le pire, une boucle se forme. Quelques mesures permettent de minimiser les risques d'erreurs (ne filtrer les VLANs que sur les commutateurs de bordure, et les laisser tous passer sur les autres), mais lorsque le nombre d'anneaux augmente, il paraît relativement périlleux de maintenir manuellement les configurations ou de le faire sans vérification. Malgré cela, le protocole est stable, permet une convergence très rapide même avec du matériel hétérogène, et l'insertion de nouveaux équipements est assez aisée.

## 6 Deux conséquences

### 6.1 La résilience et la supervision

Dans tout réseau, mettre en place de la supervision est plus que nécessaire. Toutefois, dans un réseau non résilient, la perte d'un lien ou d'un équipement se voit forcément. La supervision est un outil extrêmement important pour anticiper ou détecter au plus tôt un problème, mais elle reste un choix. A l'inverse, dans un réseau résilient, par construction, un dysfonctionnement peut passer complètement inaperçu. Pourtant, il faut pouvoir le corriger au plus tôt avant qu'un éventuel deuxième problème ne vienne bloquer le fonctionnement. Par conséquent, il ne peut pas y avoir de résilience efficace sans supervision. C'est certainement une des conditions les plus importantes pour que les administrateurs réseau et les utilisateurs aient une confiance justifiée dans l'infrastructure mise en place.

Dans notre cas, une simple interrogation SNMP toutes les minutes sur chacun des master (via Zabbix) permet de confirmer qu'on est bien en situation de fonctionnement normal.

## 6.2 La relativité de la simplicité

Les protocoles utilisés sont relativement simples à mettre en œuvre. Cependant, il faut configurer soigneusement les équipements d'un anneau, en maintenant une cohérence d'ensemble, sous peine de provoquer des dysfonctionnements et d'aller à l'encontre de l'objectif initial. Lorsque ces configurations doivent être déclinées de nombreuses fois, cela nécessite une rigueur et un formalisme accrus, ce qui peut être perçu comme plus de contraintes et de complexité.

De notre point de vue, la résilience est importante, de par les nombreux avantages qu'elle apporte. Nous avons donc cherché à voir dans quelle mesure l'opération de configuration pouvait être facilitée, pour diminuer les risques d'erreurs.

Deux approches peuvent être envisagées : la configuration automatique ou la configuration manuelle avec vérification de cohérence.

Côté configuration automatique, nous nous sommes plus particulièrement intéressés à Puppet<sup>1</sup>. Puppet est un outil de gestion centralisée de configuration, initialement de serveurs/ordinateurs. Depuis la version 2.7, il permet de réaliser quelques actions sur des équipements réseau, dont la création de VLANs et la configuration d'interfaces. Le concept est très intéressant puisqu'on peut pousser les commandes nécessaires sur les équipements. Puppet dispose d'une architecture modulaire et facilement extensible qui permet d'appliquer ces éléments de configuration à différents systèmes au travers de code spécifique à ces systèmes : les « providers ». Aujourd'hui, seul un « provider » pour les équipements Cisco existe, et nous avons d'ailleurs pu l'utiliser très facilement. Pour peu que les « providers » pour ses matériels existent (chacun peut potentiellement en développer), on peut gérer aisément un réseau hétérogène.

Nous explorons également l'aspect vérification de cohérence. Dans ce cas-là, le principe de base est que la configuration se fait toujours directement sur les équipements concernés et un processus régulier vient copier les configurations et les comparer. Nous n'avons à l'heure actuelle pas trouvé d'outil dédié à cela. Par contre, nous avons commencé à explorer l'option de combiner un outil tel que Rancid<sup>2</sup>, qui va récupérer les configurations des différents équipements réseaux, à des développements appropriés qui seraient capables d'extraire les lignes de configuration des différents équipements et de les confronter pour en signaler les éventuelles incohérences.

Nous testons activement les deux possibilités. La vérification de cohérence est certainement la plus rapide à mettre en œuvre, mais la configuration automatique a un potentiel plus grand, même si elle implique un changement d'habitudes.

## 7 Conclusion

Le réseau est une ressource de plus en plus importante, qui, au fur et à mesure de l'évolution des utilisations, supporte de moins en moins les interruptions. Or, l'évolution d'un réseau fait que de telles coupures sont inévitables. Mettre en place de la résilience, c'est décorrélater autant que possible l'utilisation du réseau qu'en font les utilisateurs des actions de maintenance préventive et curative, et permettre à chacun de travailler plus sereinement. Les architectures en boucle sont un moyen d'atteindre cet objectif à moindre coût. Après un exposé de diverses possibilités, nous avons montré un exemple concret de mise en œuvre d'une telle architecture dans un réseau de campus. Nous avons vu que cela impliquait inévitablement des changements dans la façon de gérer son réseau, mais nous espérons avoir montré que cela permettait aussi et surtout de ramener aussi bien aux utilisateurs qu'aux administrateurs réseau une certaine sérénité, bienvenue face aux évolutions et au renforcement de la criticité des réseaux.

## Bibliographie

- [1] TutoJRES n°9, Haute disponibilité des réseaux, février 2009.
- [2] TutoJRES n°10, Haute disponibilité des services, avril 2009.
- [3] Scott Hogg. 9 common Spanning Tree Mistakes, Network World, Février 2013. cf <http://www.networkworld.com/community/blog/9-common-spanning-tree-mistakes>
- [4] H3C. RRPP Technology White Paper, 2008. cf [http://www.h3c.com/portal/Products\\_\\_\\_Solutions/Technology/LAN/Technology\\_White\\_Paper/200810/618495\\_57\\_0.htm](http://www.h3c.com/portal/Products___Solutions/Technology/LAN/Technology_White_Paper/200810/618495_57_0.htm)

---

1. <http://puppetlabs.com/puppet/puppet-open-source>

2. <http://www.shrubbery.net/rancid/>