Osiris 4, un réseau métropolitain EVPN VXLAN

Sébastien Boggia, Jean Benoit

Le nouveau réseau métropolitain Osiris pour l'E/R à Strasbourg

Sébastien Boggia
Direction du Numérique
Université de Strasbourg

Jean Benoit
Direction du Numérique
Université de Strasbourg

Résumé

Osiris est le réseau métropolitain Strasbourgeois de l'enseignement supérieur et de la recherche. Jusqu'à peu, Osiris s'appuyait sur une architecture classique de transport de VLAN et de routage centralisé. Pour sa quatrième version, nous avions expliqué aux JRES 2019 notre choix d'apporter une véritable rupture technologique. Notre démarche a consisté à étudier les différentes technologies de réseaux d'overlay. Nous avons finalement opté pour EVPN/VXLAN. C'est une technologie à notre connaissance encore peu déployée sur les réseaux de campus dans l'enseignement supérieur et la recherche.

Osiris 4 consiste en une fabric de 150 équipements délivrant des services de niveau 2 et 3 jusqu'en entrée de bâtiment. Nous évoquerons son architecture et son fonctionnement, et nous présenterons la topologie choisie.

Après une phase préparatoire, le nouveau réseau a été déployé très rapidement, en 4 mois, tout en cohabitant avec le réseau existant. Nous expliquerons quelles stratégies et quels outils ont été utilisés pour atteindre les objectifs fixés.

Puis nous ferons un retour sur l'exploitation d'Osiris et nous parlerons des gains constatés, notamment au niveau des problèmes récurrents sur les versions précédentes d'Osiris.

Et enfin, nous ouvrirons sur quelques perspectives d'évolutions.

Mots-clefs

Réseau métropolitain, EVPN VXLAN, 100Gb/s, migration

1 Introduction

Osiris est le réseau métropolitain de l'enseignement supérieur et de la recherche de l'agglomération de Strasbourg. Il est opéré par la Direction du Numérique (DNUM) de l'Université de Strasbourg et apporte la connectivité réseau à 17 établissements partenaires.

Osiris, après plus de 10 ans de bons et loyaux services dans sa troisième version, a récemment évolué vers ***Osiris 4*** dans le cadre d'une refonte complète.

Aujourd'hui, *Osiris 4* s'étend sur 4 principaux campus et interconnecte 125 bâtiments. Il bénéficie de sa propre infrastructure optique utilisée pour le raccordement de plus de 95% des bâtiments.

Osiris 3 s'appuyait sur une architecture classique de transport de VLAN et de routage centralisé. Nous avons décidé d'apporter à *Osiris 4* une véritable rupture technologique par rapport à son prédécesseur en établissant une couche d'abstraction entre le réseau de transport (*underlay*) et les services (*overlay*).

Nous avions expliqué sous la forme d'un poster lors des JRES 2019 [1] à Dijon les raisons de cette rupture et notre démarche qui a consisté à étudier les différents produits et technologies du marché (*MPLS*, *EVPN/VXLAN*, *LISP/VXLAN*, *SBP*) pour trouver la solution la plus adaptée à nos besoins et à nos contraintes techniques et financières. En raison d'un rebondissement de dernière minute pendant l'écriture de l'article nous n'avions pu dévoiler notre orientation finale vers *EVPN/VXLAN* qu'une fois sur place à Dijon. Nous recommandons la lecture de cet article avant d'aller plus loin.

2 Contexte et objectifs du projet Osiris 4

Le réseau *Osiris* est renouvelé intégralement tous les 8 à 10 ans. Nous avons pu démarrer le projet *Osiris 4* car le réseau était amorti et les provisions suffisantes pour procéder à son renouvellement. Par ailleurs, il était devenu urgent de pallier les limitations de l'ancienne architecture. Enfin, le nouveau *datacenter* de l'université ayant été livré, il fallait permettre aux composantes d'y héberger leurs serveurs et leurs applications.

Chaque nouvelle version du réseau *Osiris* est financée grâce aux provisions réalisées par les partenaires *Osiris* tout au long de la durée de vie du réseau précédent. Ces provisions d'un montant d'1 million € HT ont constitué notre budget pour l'acquisition des équipements *Osiris 4*.

Non seulement le moment était venu de changer le réseau au niveau comptable, mais aussi parce que les technologies avaient évolué et qu'il était souhaitable d'aller vers une architecture plus robuste et plus facile à maintenir [1].

En l'absence de solution d'hébergement central, les composantes avaient chacune leur propres locaux contenant des serveurs. Il y avait une volonté de leur part de mettre ces ressources dans un endroit plus adapté offrant de meilleures garanties. Mais cela nécessite d'avoir des interconnexions entre le bâtiment et le *datacenter* offrant de débits très importants et un haut niveau de fiabilité.

Un autre aspect important est que le réseau *Osiris* repose sur une infrastructure optique en propre densément maillée qui nous apporte de nombreuses possibilités en matière de débit et de redondance.

Tous ces éléments nous ont conduit à fixer les objectifs qui suivent.

- Le raccordement des bâtiments standard en 10 Gb/s (à ce jour 92% des bâtiments). Cela est rendu possible par le faible surcoût, rapporté au coût global du projet (2%), des optiques 10Gb/s par rapport au 1Gb/s. Ce débit permet notamment de lever tout frein pour une migration de ressources dans le *datacenter* et de garantir une équité entre les composantes.
- Le double attachement des bâtiments (à ce jour 60% des bâtiments). La proportion de bâtiments double attachés évolue au fil des travaux réalisés sur le réseau optique.
- Un coeur de réseau résilient et de grande capacité (100 Gb/s).
- Une offre de services de niveau 2 et niveau 3 homogène en tout point du réseau : double pile IPv4/IPv6, L3VPN, transport des services de niveau 2 (très répandu sur *Osiris*).
- Une exploitation du réseau simplifiée grâce à des référentiels et des outils d'automatisation.

- Une limite de responsabilité entre le cœur de réseau et celui du bâtiment. Le **CE**[1] matérialise cette limite. Ainsi, nous conservons la maîtrise de la liaison entre les locaux de concentration et les bâtiments.

3 Choix de la solution

Dans le projet *Osiris 4*, nous nous sommes tournés très rapidement vers les technologies de réseau d'*overlay*. Nous étions conscients que celles-ci proposaient des solutions aux problématiques que nous avions sur *Osiris 3* pour :

- améliorer la fiabilité globale du réseau avec un trafic niveau 2 très présent,
- simplifier les configurations des services et des mécanismes de redondance,
- répartir la charge sur tous les liens du réseau,
- standardiser et automatiser les configurations offertes aux utilisateurs.

L'obsolescence des équipements *Osiris 3* nous offrait l'opportunité de tout remettre à plat et renouveler l'ensemble des matériels, y compris les *CE*, pour créer une *fabric*[2] de 150 équipements délivrant des services de niveau 2 et 3 jusqu'en entrée de bâtiment.

La possibilité de livrer les services sur le *CE* a été une contrainte majeure pendant toute la durée du projet. Nous voulions aussi disposer d'une densité de ports 10 Gb/s en entrée de bâtiment suffisante pour permettre un double attachement du commutateur tout en proposant plusieurs ports 10 Gb/s vers les réseaux des composantes. Cependant, les équipements répondant à ces besoins étaient annoncés par les constructeurs mais n'étaient souvent pas encore disponibles, ce qui a engendré des retards importants dans le projet. Les équipements répondant à ces besoins et entrant dans l'enveloppe budgétaire n'existaient pas initialement.

Deux technologies se sont trouvées en position de finalistes pour *Osiris 4* : *SPBm* [2] et *EVPN/VXLAN* [3]. Ces deux technologies répondent à notre cahier des charges fonctionnel. Au moment des études menées sur les différentes technologies, *EVPN/VXLAN* était très orienté réseaux de datacenter et à notre connaissance encore peu déployé sur les réseaux de campus. SPBm était particulièrement adapté aux réseaux de campus et permettait en outre une prise en main rapide avec une configuration et un mode de fonctionnement moins complexe qu' *EVPN/VXLAN*.

Notre choix final s'est malgré cela porté sur *EVPN/VXLAN* avec le constructeur Arista, et ceci pour plusieurs raisons.

Tout d'abord, les équipements répondant à nos besoins étaient disponibles, et l'ensemble des fonctionnalités demandées pour fournir les services *Osiris* était déjà implémentées.

Ensuite, les négociations financières entamées avec les équipementiers finalistes ont abouti à des prix sensiblement comparables. Le *CE* Arista bénéficie en outre d'une grande densité de port pour un prix très compétitif apportant plus de souplesse pour le raccordement des bâtiments.

Le réseau récemment déployé dans le *datacenter* de l'Université de Strasbourg s'appuyait sur la même technologie et le même constructeur. Nous avions donc un bon retour d'expérience avec la *fabric* de notre *datacenter* qui affichait une très bonne fiabilité. Et la possibilité de mutualiser avec *Osiris 4* des outils et le système d'information mis en place dans le cadre du d*atacenter* apportait un gain de temps non négligeable

1 CE : Commutateur d'Extrémité, marque la limite entre le backbone Osiris et le réseau du bâtiment

2 Network Fabric : désigne les réseaux utilisant les technologies d'overlay présentant une abstraction entre le réseau de transport et les services

Enfin, *EVPN/VXLAN* est devenu un standard proposé par les constructeurs leaders du marché, ce qui nous rassure dans la pérennité de notre choix.

4 Architecture Osiris 4

La topologie du réseau *Osiris 4* est basée sur une architecture à 3 niveaux (Figure 1) :

- ***leaf***, commutateur d'entrée de bâtiment (*CE*) livrant les services aux composantes, double-attaché au cœur,
- ***spine***, commutateur de concentration des bâtiments au niveau d'un campus, avec 2 *spines* par campus dans 2 *PoP*[3] (point de présence) distincts,
- ***super-spine***, commutateur de concentration des spines pour concentrer l'ensemble des campus et interconnecter *Osiris 4* au *datacenter* et à Renater[4].

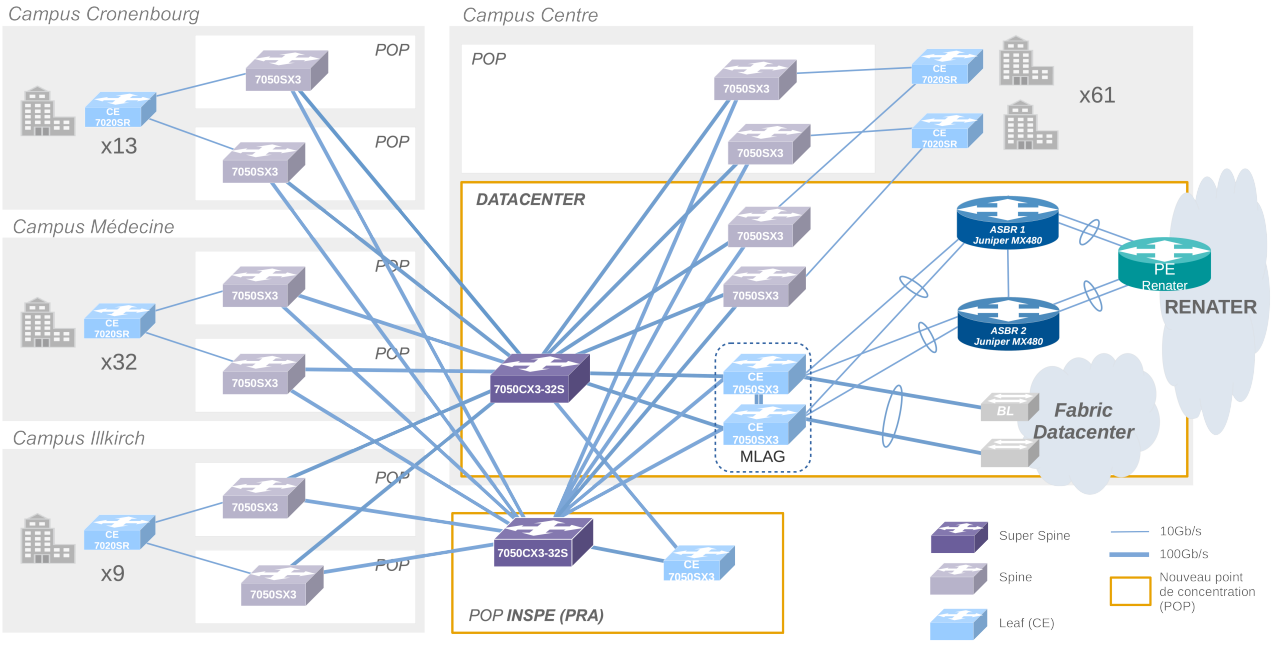

Figure 1: Architecture réseau Osiris 4

Cette architecture à 3 niveaux se justifie par la capillarité du réseau optique sous-jacent et la distance entre les différents campus (parfois plus de 10 km), qui ne permet pas de centraliser les connexions de tous les bâtiments en un point unique. Toutefois, la densité et le maillage du réseau optique nous permet au moins de faire un double raccordement systématique de chaque *spine* vers les deux *super-spines*.

Chaque *CE* est connecté au *spine* avec 2 liens 10 Gb/s lorsque la disponibilité en fibre optique est suffisante. Les interconnexions entre les *spines* et les *super-spines* sont des liens 100 Gb/S.

Cet agencement permet une forte redondance et une grande efficacité de répartition du trafic : chaque paquet peut emprunter plusieurs chemins de même longueur, ce qui permet d'atteindre jusqu'à 400 Gb/s de capacité par campus. Et même en cas de rupture d'un lien ou de panne d'un

3 PoP : Point of presence. Point d'interconnexion d'équipements réseau, ici les noeuds de concentration des liaisons vers les bâtiments Osiris.

4 RENATER : Réseau National de élécommunications pour la Technologie, l'Enseignement et la Recherche

équipement de cœur, il n'y a pas d'allongement de la longueur des chemins. Nous estimons que ce dimensionnement sera suffisant pour tenir jusqu'à la fin d'Osiris 4, d'ici 8 à 10 ans.

Les équipements choisis sont les suivants :

- *super-spine* : Arista 7050CX3 (32 ports 100Gb/s)
- *spine* : Arista 7050SX3 (8 ports 100Gb/s + 48 ports (10/25 Gb/s)
- *leaf* (CE) : Arista 7020SR (2 ports 100 Gbs/s + 24 ports 10Gb/s)

Il s'agit de commutateurs 1U qui offrent une densité de ports 10G/100G suffisante pour interconnecter entre eux les campus et les bâtiments.

Le campus *Centre* avec près de 70 bâtiments double-attachés exige un nombre de ports qui aurait pu nécessiter un déploiement de châssis pour les *spines*. Or les châssis coûtent plus cher, nécessitent plus de manutention et consomment plus d'énergie. Il nous a paru plus adapté de mettre 2 *spines* 1U dans chaque PoP du campus *Centre* pour apporter la densité nécessaire et rester sur une gamme d'équipements *spine* homogène.

L'homogénéité facilitant l'exploitation du réseau, nous avons aussi choisi un seul modèle d'équipement pour les *CE (*sauf *CE datacenter)*. En outre, ce modèle dispose de deux ports 100Gb/s qui permettrons, si nécessaire, de faire évoluer le connectivité vers le coeur de réseau.

Comme nous l'avons évoqué plus haut, le réseau *Osiris 4* est une *fabric EVPN/VXLAN (Figure 2)* qui permet de virtualiser des services réseaux avec une couche d'abstraction entre un réseau de transport, *underlay network*, et un réseau de service, *overlay network*.

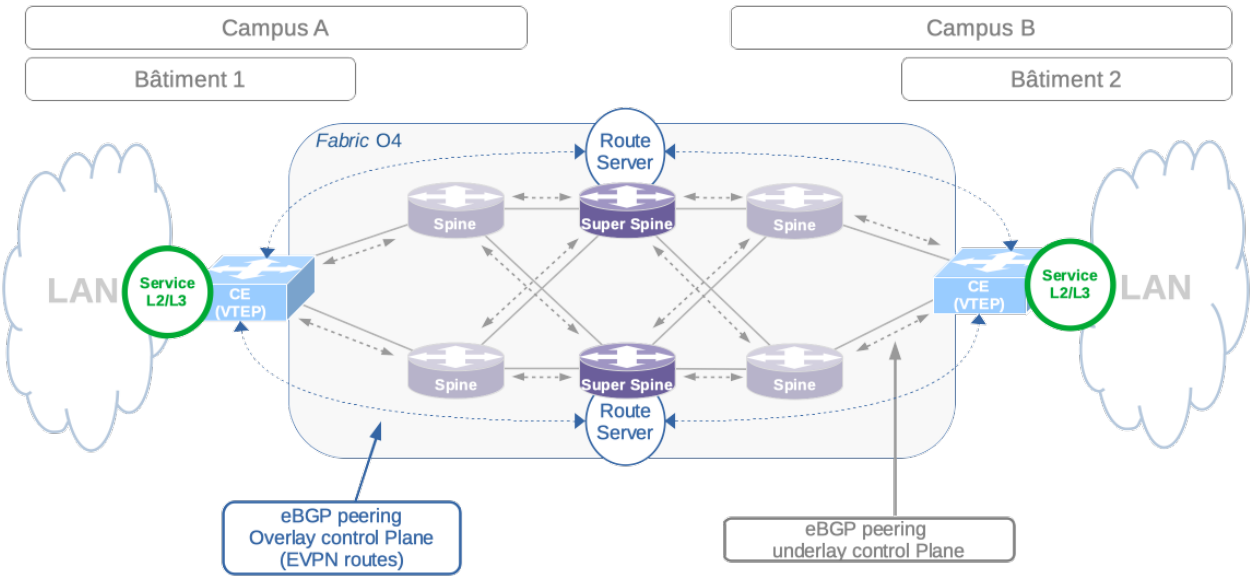

Figure 2: Fabric EVPN/VXLAN Osiris 4

Le rôle de l'*underlay* est de rendre accessible entre eux les différents *CE*. Chaque *CE* est un point de terminaison de services que l'on nomme *VTEP*[5] dans la terminologie *VXLAN*. C'est le *VTEP* qui encapsule et décapsule les paquets des réseaux utilisateurs et les transporte via l'*underlay* d'une extrémité à l'autre de la *fabric*.

5 VTEP : VXLAN Tunnel End Point

Le réseau *underlay* est basé sur du routage IP. Il nécessite des réseaux d'interconnexion entre les équipements et des *IP* de *loopback* pour identifier les *VTEP*. La diffusion des routes pour assurer la connectivité réseau peut être effectuée par un protocole comme *OSPF* ou *IS-IS*. Dans notre cas, nous avons choisi *eBGP* tel que conseillé par le constructeur.

La configuration de *BGP* peut paraitre fastidieuse (attribution des numéros d'*AS*, adresses IP pour les *peering*, etc.) mais elle a été grandement simplifiée ici. Les *spines* acceptent les *peering BGP* des *CE* dynamiquement et la quasi-totalité des éléments de configuration sont générés automatiquement à l'aide d'outils dont nous parlerons plus loin.

En *overlay*, *EVPN* utilise *BGP* pour distribuer les routes des services de niveau 2 ou 3. La signalisation s'appuie sur des r*oute-servers* hébergés sur les *super-spines*. Chaque *CE* monte un *peering* (à partir d'une *loopback*) vers ces *route-servers* et y annonce les réseaux IP et les adresses MAC associés aux services. Les *route-servers* vont ré-annoncer ces routes *EVPN* vers les autres *CE*. La destination des routes (next-hop) correspond à la *loopback* associée à la *VTEP*.

Le routage *ECMP*[6] est activé dans *BGP*. Il va permettre de répartir le trafic sur tous les chemins de même longueur. La topologie d'*Osiris 4* offre jusqu'à 4 chemins possibles entre 2 extrémités du réseau.

En bordure de réseau, deux *CE* redondants (*Figure 1*) réalisent l'interconnexion avec le *datacenter* et *Renater*. Ce point sensible du réseau doit résister à une panne d'un équipement ou d'un lien. Pour ce faire, nous utilisons des agrégats répartis sur les deux *CE* (MLAG *[4]*). Ces *CE* sont du même modèle que les *spines* pour disposer de suffisamment de ports 100 Gb/s. Les fabrics *datacenter* et *Osiris* étant indépendantes, la continuité des services niveau 2 et niveau 3 entre elles se fait en prolongeant des VLAN.

5 Migration vers Osiris 4

La migration vers *Osiris 4* s'est étalée sur plusieurs mois et a demandé un travail préparatoire important. Les différentes étapes ont été :

- la préparation de l'infrastructure optique,
- le déploiement d'un réseau hors-bande de management,
- l'adaptation des outils de déploiement,
- le déploiement du cœur 100G (*spines* et *super-spines*),
- l'élaboration de la stratégie de migration des sites,
- le travail de préparation avec les correspondants informatiques de sites,
- la migration des 120 sites.

5.1 Infrastructure optique

Il a tout d'abord fallu procéder à des modifications conséquentes sur le réseau de fibres optiques très en amont du déploiement d'O*siris 4*, dans le contexte de la construction du nouveau *datacenter*.

Le *datacenter* ainsi que le bâtiment de l'*INSPE* (désigné pour héberger le PRA) allaient devenir les deux nouveaux principaux nœuds de concentration *(Figure 1)*. Le réseau optique a été pensé pour

6 ECMP : Equal Cost Multi Path

que chaque *PoP* soit connecté aux nœuds de concentration par deux chemins optiques totalement différents. Ainsi de nouvelles liaisons ont été construites depuis le nouveau *datacenter* :

- vers tous les bâtiments du campus Centre,
- vers les nœuds de concentration des autres campus (*PoP*).

De même, une deuxième adduction optique a été réalisée au bâtiment de l'INSPE pour y amener les liaisons allant vers les *PoP* sans utiliser de chemin commun avec le d*atacenter*.

Face à l'ampleur des modifications, les travaux sur l'infrastructure optique ont dû être démarrés avant même que les matériels actifs et la technologie utilisés ne soient connus. Le réseau optique a ainsi été provisionné pour s'adapter aux différentes hypothèses de topologie *Osiris 4*.

Enfin, du côté des bâtiments, les travaux réalisés ont permis d'augmenter la proportion de bâtiments double attachés aux *PoP* des campus. Pour supporter des débits de 10Gb/s, plusieurs fibres optiques multimodes anciennes[7], ont été remplacées par de la monomode.

5.2 Réseau hors bande

Un réseau de management *Out Of Band*[8] a été déployé parallèlement à Osiris 4. Il est architecturé autour d'un anneau de commutateurs déployés à travers les *PoP sur une infrastructure optique indépendante* et il raccorde chaque équipement de la *fabric* via leur interface ethernet de management.

Le réseau hors-bande (*OOB*) a été un prérequis pour la migration. Il permet la configuration initiale, l'accès au management de tous les équipements, le monitoring et la métrologie. Le réseau hors bande existait auparavant sur *Osiris 3* mais uniquement pour offrir une solution de secours pour l'accès au management des équipements de concentration. Il s'est montré très utile lors d'incidents ou pour des opérations de maintenance. À la mise en place du réseau *datacenter*, l'accès hors-bande a été généralisé dès le départ pour le déploiement de l'ensemble des équipements. Nous avons repris le même principe pour *Osiris 4* pour tous les équipements de la *fabric*.

5.3 Outils de déploiement

Pour mutualiser les efforts, nous avons décidé de réutiliser les outils de gestion développés pour le réseau *datacenter*. Ils ont nécessité une adaptation pour *Osiris 4*.

Ces outils s'articulent autour d'un référentiel réseau, nommé *R2DC*, qui va contenir l'ensemble des éléments variables des configurations des équipements : noms, numéros de série, interfaces, adressage IP, numéros d'AS, ainsi que les services de niveau 2 et niveau 3 (VNI, route targets, …).

La configuration initiale d'un équipement est déployée via *ZTP*[9] selon un modèle type (*super-spine*, *spine* ou *leaf*) et un mécanisme qui génère automatiquement les éléments *variables* de la configuration en interagissant avec le référentiel. Lorsque le processus *ZTP* est terminé, l'équipement est opérationnel : il est intégré un réseau *underlay* de la *fabric*, ses peerings e*BGP* avec les *route servers* sont établis pour configurer les services en *overlay*.

Un autre outil, *Odile*, est utilisé pour configurer les services en *overlay*. Il interagit avec le référentiel pour le provisionning des éléments variables dans la configuration des services.

7 Fibres optiques multimodes OM1 n'offrant qu'une portée de quelques dizaines de mètres en 10Gb/s

8 Out Of Band (OOB), le réseau OOB ne transporte pas de trafic utilisateur à l'inverse de réseau In Band

9 ZTP (Zero Touch Provisioning) : permet une configuration initiale de l'équipement entièrement automatisée lors de son déploiement

Le but de ces outils est de ne pas intervenir sur les équipements en *CLI*[10] pour réaliser des configurations et ainsi limiter les erreurs. Ils ont été décrits dans l'article concernant le nouveau *datacenter* de l'Unistra aux JRES 2019 [5].

5.4 Déploiement du coeur 100G

Les trois premières étapes décrites ci-dessus ayant été réalisées, nous avons pu déployer le cœur 100G d'*Osiris 4* (*super-spines* et *spines*). C'était une opération simple à réaliser car la partie optique était prête et la configuration des équipements était entièrement automatisée. Il a simplement fallu effectuer la mise en place dans un ordre précis. D'abord les *super-spines*, ensuite les *spines*, puis les deux *CE avec les* MLAG prévus pour interconnecter *Osiris 4* à *Renater* et au *datacenter*. L'ensemble de ce déploiement a été effectué en environ une semaine (*Figure 1)*.

5.5 Stratégie de migration des sites

Lorsque nous avons élaboré la stratégie de migration des 120 sites, il paraissait évident que celle-ci s'étalerait sur plusieurs mois. Or, le temps de basculer les sites d'un réseau à l'autre, *Osiris 3* et *Osiris 4* devaient fonctionner en parallèle. Nous avons ainsi interconnecté sous forme d'un agrégat 2x10Gb/s (*MLAG*) les deux *CE* situés au *datacenter* à l'un des principaux commutateurs de concentration d'*Osiris 3 (Figure 3).* Si cette interconnexion avait pour avantage d'être simple et d'apporter le débit nécessaire pour éviter les congestions, elle avait pour inconvénient de transformer en *SPOF*[11] le commutateur *Osiris 3* en fin de vie. Avec cette architecture fragile, la migration devait être très rapide. Nous avons décidé de la réaliser à un rythme très soutenu en moins de 6 mois.

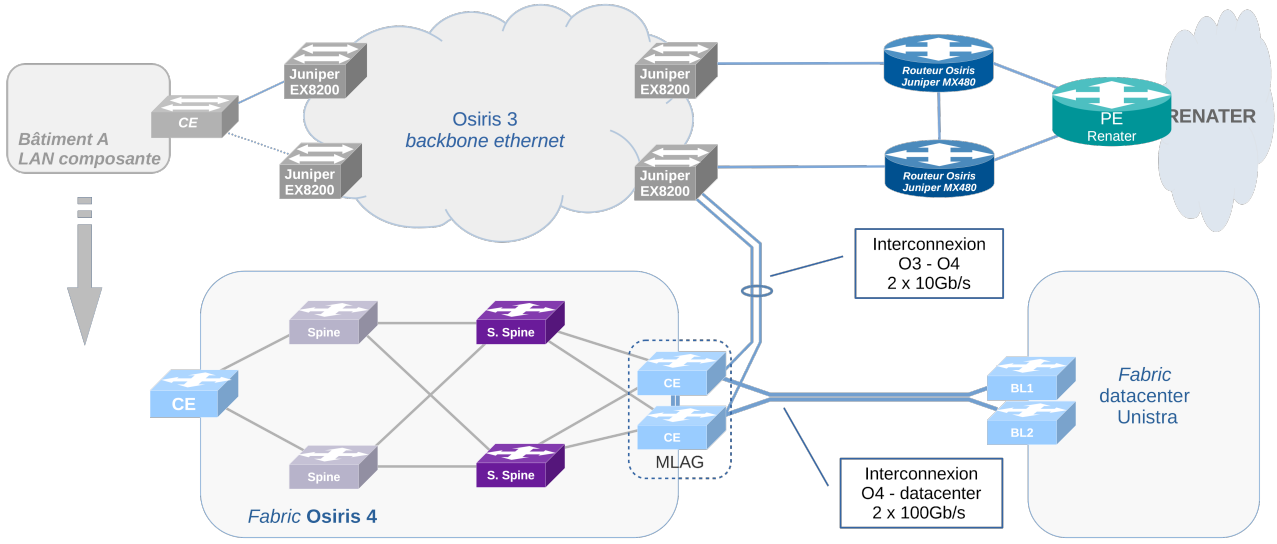

Figure 3: Migration, pont Osiris 3 - Osiris 4

Sur *Osiris 3*, le routage des réseaux des composantes était centralisé sur deux routeurs *Juniper MX480*. Le protocole *VRRP* assure la redondance en mode actif-passif. Pour simplifier et accélérer la migration des sites vers *Osiris 4*, nous avons décidé de ne pas déplacer les points de routage de ces réseaux publics sur les *CE Osiris 4* mais de conserver les MX480 d'*Osiris 3* dans un premier temps, quitte à migrer le routage plus tard. Ainsi la migration, consistait à prolonger les réseaux ethernet des *CE* jusqu'au point de routage sous forme de service de niveau 2 (*Figure 4*).

10 CLI : Command Line Interface

11 SPOF : Single Point Of Failure ou point de défaillance unique dont le reste du système d'information est dépendant. Il n'est pas redondé.

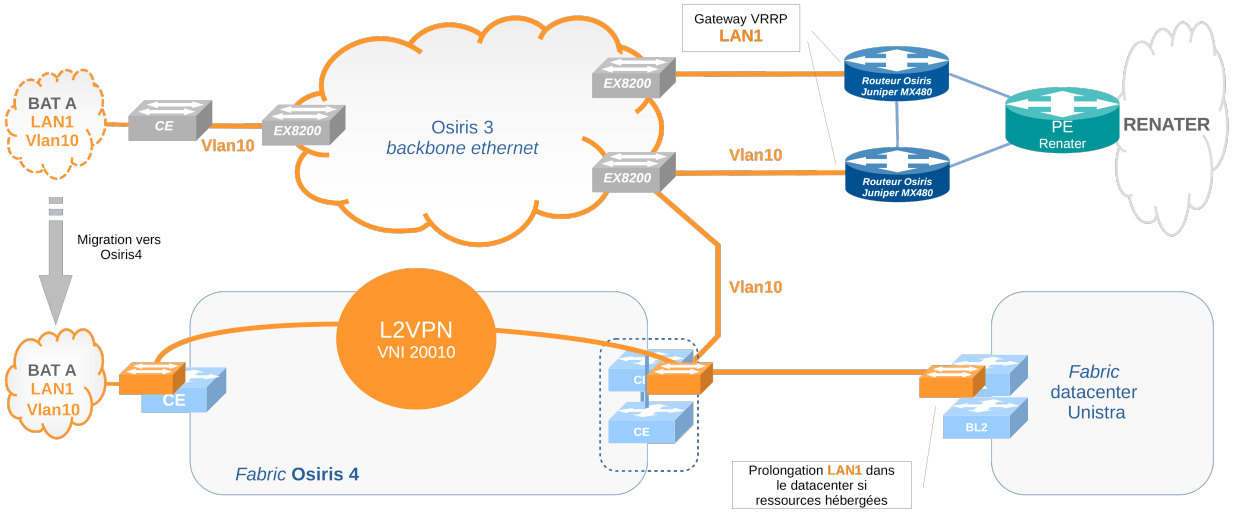

Figure 4: Migration service de niveau 2 (L2VPN)

Seuls le routage des *VRF lite* [6] d'*Osiris 3* a été migré sous forme de niveau 3 dans *Osiris 4*. Dans cette opération, pour chaque *VRF*, des interconnexions IP ont été créées pour échanger les routes entre *Osiris 3*, *Osiris 4* et la *fabric* du d*atacenter (Figure 5)*.

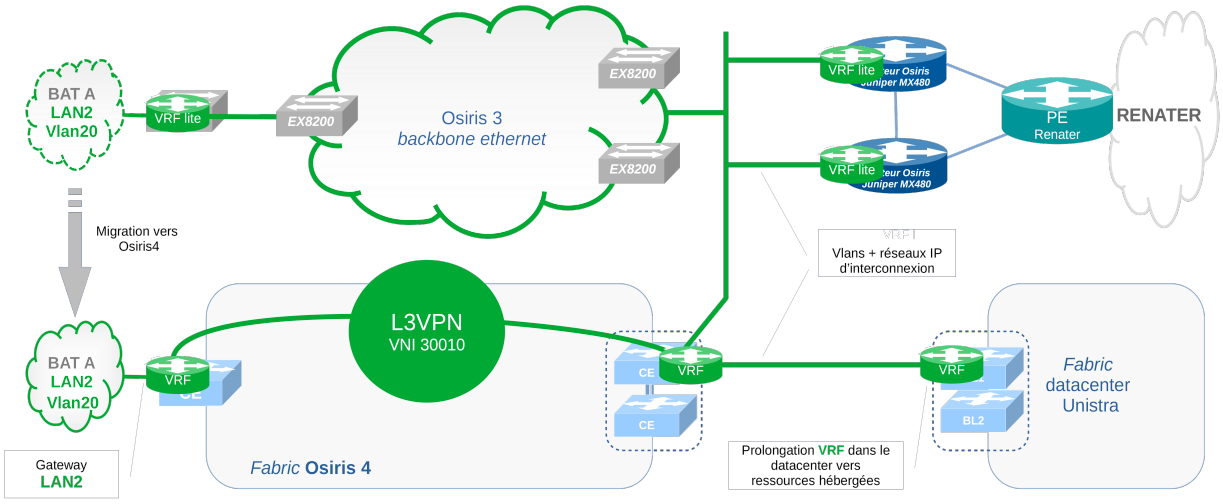

Figure 5: Migration service de niveau 3 (L3VPN)

Jusqu'à présent, la livraison du service dans un bâtiment était faite sur des port ethernet distincts. Pour la migration *Osiris 4*, nous avons défini une nouvelle spécification d'interconnexion :

- 1 port 10Gb/s par réseau de composante,
- tous les services livrés sur ce port (tags de *VLAN 802.1Q*).

Cela a nécessité des modifications côté réseaux des composantes (passage de N ports accès à 1 port *trunk 802.1q*) et des interactions avec les correspondants locaux. Ensuite, nous avons formalisé la procédure de migration d'un site et développé les outils avant de commencer les déploiements.

5.6 Processus de déploiement des sites

5.6.1 Étape 1 : information préalable

Environ un an avant le début de la migration des sites vers *Osiris 4*, un document intitulé *DSCS*[12] a été diffusé à l'ensemble des correspondants OSIRIS.

Ce document définit les prérequis environnementaux minimaux à respecter pour la migration.

- nombre de U nécessaires et profondeur minimale de la baie pour accueillir les équipements : un *CE* et un commutateur Hors Bande,
- température ambiante maximale dans le local technique,
- alimentations et puissance électrique nécessaires.

Il a permis d'identifier et de corriger les problèmes environnementaux suffisamment en amont.

5.6.2 Étape 2 : diffusion du questionnaire

Quelques semaines avant le début des opérations, nous avons lancé un script pour auditer les configurations actives sur chacun des *CE Osiris 3* à migrer et pour générer un questionnaire au format *CSV* envoyé au correspondant informatique du site *(Figure 6)*. Celui-ci devait remplir le fichier et nous le retourner en répondant aux questions suivantes :

- Les réseaux présents sont-ils toujours utilisés ?
- Est-il possible d'interconnecter le réseau de la composante selon les nouvelles spécifications ?
- Les conditions environnementales sont-elles conformes au DSCS ?

questionnaire_ics-ce1

DESCRIPTION DU RESEAU	TAG 802.1Q	RESEAU	INTERFACES ACTUELLES	RESEAU A ACTIVER SUR OSIRIS 4 (O/N)	TAG 802.1Q DESIRE VERS LAN
Bornes Wi-Fi controlees	914	10.2.72.0/24	ge-0/0/0	Oui	914
Inter UMR ECPM - IPCMS - ICS	2010		ge-0/0/6	Oui	2010
rch unistra ics	24	130.79.210.224/28	ge-0/0/0	Oui	2
INTERCONNEXION ENTRE LE RESEAU DU BATIMENT ET OSIRIS 4	(O/N)	REMARQUE			
Port 10G transportant les reseaux sur des tag 802.1q	Oui				
Raccordement DAC	Oui				
ENVIRONNEMENT	(O/N)	REMARQUE			
Profondeur minimale de 550 mm	Oui				
2U disponibles dans l'armoire	Oui				
2 prises electriques	Oui				
Puissance de 300W disponible	Oui				
Espace de 10cm en facade pour les fibres optiques	Oui				
Armoire aeree avec temperature ambiante de 40 degre maximum	Oui				

Modification du TAG 802.1q souhaité par la composante

La composante confirme qu'une interconnexion avec le CE O4 en 802.1q et 10G DAC est possible

La composante confirme que les pré-requis environnementaux minimum sont respectés

Figure 6: Questionnaire de migration

Cet audit nous a permis d'anticiper les difficultés et de trouver des solutions lorsque la composante n'était pas en mesure de s'interconnecter à *Osiris* selon les nouvelles spécifications. Il a aussi permis de faire du nettoyage en supprimant bon nombre de réseaux qui n'étaient plus utilisés, en se basant sur les ports encore actifs ou sur la réponse du correspondant.

12 DSCS : Document de Spécification et de Conception du Système interne à l'Université de Strasbourg détaillant les bonnes pratiques dans les locaux techniques

5.6.3 Étape 3 : intégration des réponses

Juste avant de procéder à la migration du *CE*, le fichier *CSV* modifié est repris, validé puis traité à l'aide d'un script qui produit une représentation intermédiaire simplifiée des services à provisionner. Ce script sait à partir du *CSV* faire la différence entre :

- Les réseaux qui seront routés sur les *Juniper MX480* comme c'était le cas sur *Osiris 3* sont migrés sur *Osiris 4* sous forme de services de niveau 2 (*L2VPN*).
- Les réseaux qui appartiennent à des *VRF lites* dans *Osiris 3* sont migrés sur *Osiris 4* sous forme de services de niveau 3 (*L3VPN*).

Le script crée un fichier *(Figure 7)* qui servira à générer les commandes de configuration *Odile* des services sur *Osiris 4*. L'extrême simplicité du fichier permet de modifier manuellement certains paramètres comme le nom de l'interface à laquelle associer les services (en cas de présence de plusieurs composantes dans le bâtiment).

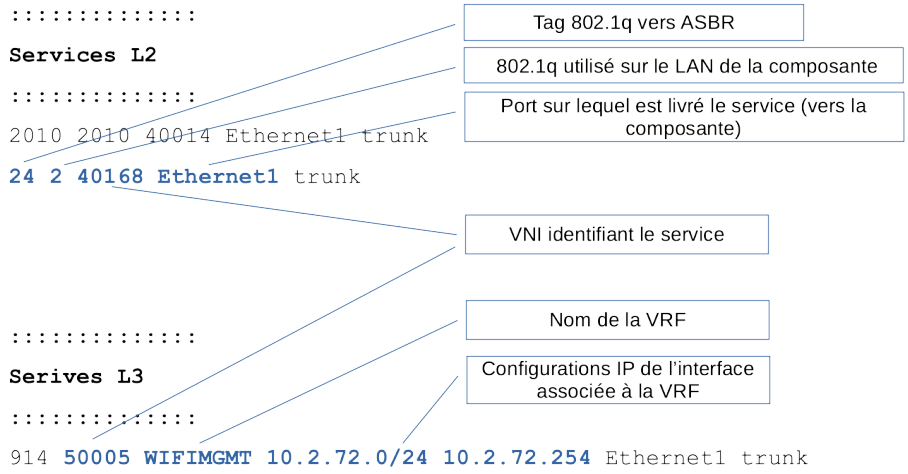

Figure 7: Fichier de génération des configuration des services

5.6.4 Étape 4 : mise en place des équipements

Le jour de la migration, on commence par le déploiement du commutateur hors bande, puis le nouveau *CE* et on met en place une console déportée (sur un Raspberry PI) connectée au réseau hors-bande *(Figure 8)*.

5.6.5 Étape 5 : connexion des fibres et démarrage des équipements

Les fibres optiques sont déconnectées de l'ancien *CE* et raccordées au *CE Osiris 4*.

Parallèlement, d'autres personnes vont rebrasser les fibres dans les locaux de concentration (*PoP*). Dès que les 2 liens connectés aux *spines* montent, le processus *ZTP* est lancé.

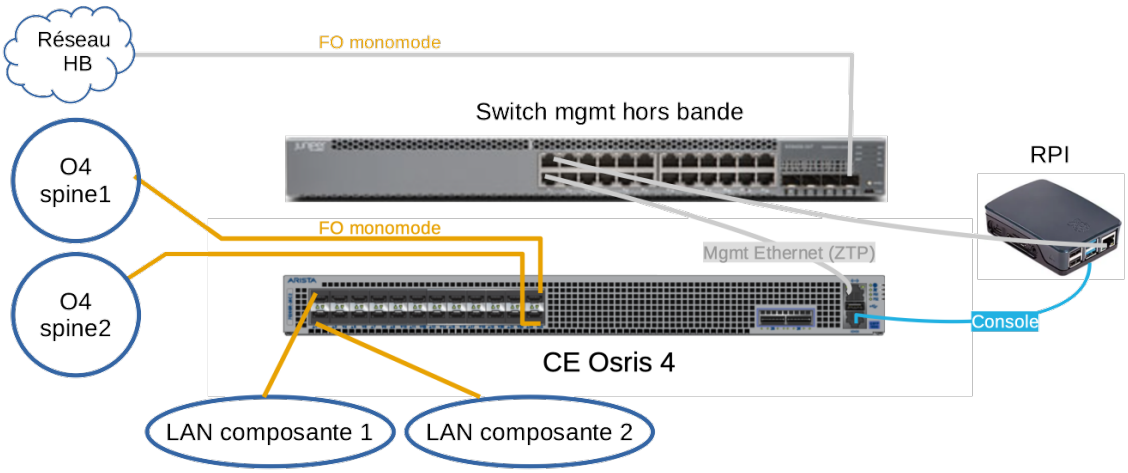

Figure 8: Installation CE et OOB

5.6.6 Étape 6 : configuration des services

Une fois le processus *ZTP* terminé, Le *CE* est intégré à la *fabric Osiris 4*. La configuration des services peut être réalisée. Les commandes *Odile* de configuration sont générées et appliquées au *CE*.

Voici un exemple de configuration du service niveau 2 (L2VPN) sur un *CE*.

```
# Configuration du port Ethernet 1 du host ics-ce1 en mode trunk
odile-cli create trunkinterface --env o4 --host 'ics-ce1' --int 'Ethernet1' --vlans none
# Création du service niveau 2 associé au VNI 40168 et au Vlan 2
odile-cli create l2service --env o4 --host ics-ce1 --vlan 2 --vni 40168
# Association du Vlan 2 au port Ethernet 1
odile-cli add vlan --env o4 --host ics-ce1 --int Ethernet1 --vlans 2
```

5.6.7 Étape 7 : connexion des composantes et tests

Le nouveau *CE Osiris 4* est opérationnel. Chaque réseau de composante est connecté sur le port ethernet correspondant. Les tests de validation de la connectivité sont réalisés.

5.7 Planification et organisation

La migration a été divisée en plusieurs phases, chaque phase se concentrant sur les sites d'un même campus. Le regroupement par campus rendait la logistique plus simple. Avant chaque phase, les collègues en charge du réseau optique se chargeaient de valider les chemins optiques et de connecter les nouveaux liens.

La migration a mobilisé l'ensemble de l'équipe et s'est déroulée en flux tendu pour tenir les délais annoncés et maintenir l'efficacité acquise au fil des déploiements.

Chaque migration s'organisait autour d'une fiche documentant tous les éléments nécessaires pour sa réalisation *(Figure 9)* : la structure des liens optiques, les informations de contact, les équipements et la connectique utilisés etc. Un autre élément important dans l'organisation était la préparation avec les correspondants de composantes. Plusieurs réunions avaient déjà eu lieu et une dernière validation avec les correspondants se faisait quelques jours avant le déploiement sur site.

Quelques jours ouvrés avant, la fiche de migration était complétée, l'ensemble du matériel était préparé, les équipes étaient constituées et des annonces étaient envoyées à l'ensemble des utilisateurs concernés.

Le rythme des migrations était soutenu : nous avons assuré quatre déploiements pour chaque journée de migration, avec deux journées de migration par semaine. Chaque jour de migration mobilisait deux équipes constituée chacune de deux personnes sur site et d'une personne dans les locaux de concentration (*PoP*). Après trois semaines de migration, le processus était bien rôdé. Néanmoins, nous avons souvent dû re-planifier des opérations en raison d'événements imprévus.

Site : LEBEL

action effectuée	Task	Sub Task	Valeur
	POP1		
☑		nom	DC
☑		numéro de lien	82
☑		SFP spine vers site prêt (type à préciser)	10G-LR Arista, dc-s1 eth27
☑		jarretière optique nécessaire	LC-LC mono
☑		préparation jarretièrage	récupérer jarretière LC-LC en place. débrancher vers
	POP2		
☑		nom	LE7
☑		numéro de lien	2113
☑		SFP vers site prêt (type à préciser)	10G-LR Arista le7-s1 eth27
☑		jarretière optique nécessaire	SC-SC Lebel, LC-SC 5m le7 mono
☑		préparation jarretièrage	**JARRETIERAGE DÉJÀ EFFECTUÉ** : Opération réalisée le 24/02/2021. Test connectivité sur le7-s1 -> OK. **Pas besoin de se déplacer sur le pop**
	Site		
☑		SFP vers POP1 prêt (type à préciser)	10G-LR
☑		SFP vers POP2 prêt (type à préciser)	10G-LR
☑		préparation jarretièrage vers POP1	récupérer jarretière en place.
☑		préparation jarretièrage vers POP2	récupérer jarretière en place
	CE		
☑		Nom	lebel-ce1
☑		N° SIFAC attribué	OUI
☑		Inventaire GLPI effectué	OUI
	Interco CE - réseau de bâtiment		
☑		Port 1 -> LAN	vers DNUM
☑		SFP vers réseau de bâtiment prêt (type à préciser)	DAC
☑		port CE 04	ethernet1
☑		type jarretière optique (si nécessaire)	N/A
☑		Port 2 -> LAN	vers FAV Chimie (JC Pont)

Figure 9: Extrait de fiche de déploiement

Une semaine type de déploiement ressemblait à ceci :

- **lundi** : préparation des migrations de mardi; mise au point avec les correspondants pour les migrations des semaines à venir.
- **mardi** : migrations de **4 sites**.
- **mercredi** : préparation des migrations du jeudi ; mise au point avec les correspondants pour les migrations des semaines à venir.
- **jeudi** : migrations de **4 sites**.
- **vendredi** : mise à jour du planning, envoi des annonces des migrations à suivre et préparation du matériel.

Finalement, nous avons migré l'ensemble des sites en **4 mois et 1 semaine**. Le dialogue avec les correspondants a été très coûteux en temps (plusieurs dizaines d'heures) mais indispensable pour que tout fonctionne le jour de la migration.

6 Retour d'expérience

Par rapport aux attentes initiales, à savoir l'amélioration des débits, de la stabilité du réseau et la simplification de l'exploitation, les objectifs ont été globalement atteints.

Le réseau a un niveau de stabilité très supérieur à *Osiris 3*. Après 7 mois d'exploitation, il n'y a pas eu d'incident.

La structure maillée du réseau et le dimensionnement des liens sont largement suffisants pour aborder les prochaines années sereinement. En outre *ECMP* répartit le trafic de façon homogène entre les différents liens de même longueur. Entre *Osiris 3* et *Osiris 4* nous sommes ainsi passés d'une redondance de liens de type active-passive à une redondance active-active.

On constate que les services de niveau 2 (*L2VPN*) ne sont pas près d'être abandonnés. En effet, la plupart des migrations de machines des composantes dans le *datacenter* est accompagnée de prolongation de services niveau 2 à travers *Osiris 4*.

Lié à cet usage abondant de service de niveau 2, sur *Osiris 3* des boucles engendrant des tempêtes de *broadcast* impactaient parfois la stabilité globale du réseau. Sur *Osiris 4*, l'impact est limité uniquement au réseau directement concerné. Par ailleurs, un mécanisme natif à EVPN de détection de *host-flapping*[13] qui met en liste noire les adresses *MAC* qui bagottent à cause d'une boucle en limite l'effet tout en protégeant le plan de contrôle des équipements.

Ce dispositif n'est toutefois pas suffisant. En cas d'amplification de trafic *BUM*[14] dans un réseau de composante, celui-ci est transmis à travers la *fabric* à tous les *CE* sur lesquels est déployé le service par un mécanisme de *Head End Replication* à partir du *CE* de la composante. Nous sommes ainsi en train de renforcer la protection contre les boucles configurant sur chaque port de *CE* à destination d'une composante une fonctionnalité appelée *Loop protection*. Ce mécanisme simple envoie un paquet spécifique vers le réseau de la composante. Si le paquet revient vers le *CE*, celui-ci considère qu'il existe une boucle sur le réseau et coupe le port.

Ces mécanismes changent l'approche des correspondants : comme le réseau victime de boucle est pénalisé, cela incite les correspondants à corriger les problèmes dans leurs réseaux.

Au niveau de l'exploitation du réseau, nous avons mis en place trois éléments importants : un réseau hors-bande généralisé jusqu'au *CE*, un processus de déploiement initial via *ZTP*, et un ensemble d'outils de gestion du réseau basé sur un référentiel. Ces trois éléments apportent les gains suivants :

- un accès robuste au management de l'ensemble des équipements,
- un déploiement et un remplacement des équipements facilités. Par exemple, le remplacement d'un *CE* défectueux se limite à la modification du numéro de série dans le référentiel *R2DC* et dans le service *ZTP* par celui du nouveau switch, puis l'installation de celui-ci sur site.
- une réduction des erreurs de configuration,
- une accélération et une simplification de la livraison du service qui ne fait plus qu'en extrémité (sur le CE), sans faire de configuration sur des équipements intermédiaires.

13 EVPN host flapping : Sur la fabric Osiris 4, le host flapping est détecté via les annoncent BGP d'une même MAC provenant de plusieurs CE.

14 BUM : Broadcast Unicast Multicast

7 Perspectives

Le réseau de fibres optiques sur lequel repose Osiris, va continuer à évoluer dans les années qui viennent. Les dernières fibres multimode qui adductent les bâtiments sont progressivement remplacées dans le cadre de travaux de rénovation pour arriver à 100% de raccordements 10Gb/s (92% actuellement). Ces travaux de rénovation sont parfois accompagnés de double attachements optiques quand cela est possible (60% de double attachements à ce jour).

Une autre évolution à apporter au réseau Osiris consistera à remplacer les 2 routeurs *Juniper MX480* interconnectés à Renater. Comme nous l'avons expliqué auparavant, ils réalisent le routage des réseaux IP publics. Pour simplifier la migration, nous avions décidé de conserver temporairement ces fonctions de routage sur ces équipements. Or, ceux-ci arrivent en fin de vie. De plus, ils ne supportent que des interfaces 10Gb/s. Nos liens 2x20Gb/s vers *Renater* sont à saturation. Nous prévoyons de supprimer ces équipements et d'interconnecter *Osiris* directement en 2x100Gb/s à *Renater* par l'intermédiaire du réseau régional *RAREST* en cours de renouvellement.

Nous sommes en train d'étudier différentes possibilités et notamment une architecture décentralisée où le routage des réseaux publics serait fait sur les *CE* dans une VRF "INTERNET". ll est aussi possible de remplacer les anciens routeurs et de conserver un routage centralisé. À ce stade, nous n'avons pas encore décidé.

8 Conclusion

Ce projet aux objectifs ambitieux s'est étalé sur plusieurs années, il nous a demandé beaucoup d'efforts, il a connu de nombreux rebondissements et les choix à faire n'ont pas toujours été faciles. Le résultat est très satisfaisant en matière de performance, de redondance, de facilité d'exploitation et d'évolutivité. Du côté du budget, entre le début de la veille et l'achat des équipements, le temps à largement joué en notre faveur pour atteindre nos objectifs techniques sans sortir de l'enveloppe attribuée en début de projet. Les objectifs fixés au début du projet ont été atteints.

9 Bibliographie

[1] JRES 2019 : Osiris 4, quelle technologie pour le MAN de Strasbourg, https://conf-ng.jres.org/2019/document_revision_5471.html?download

[2] SPB, (Shortest Path Bridging) : https://en.wikipedia.org/wiki/IEEE_802.1aq

[3] EVPN, https://en.wikipedia.org/wiki/Ethernet_VPN

Xavier Jeannin - Yann Dupont - Anthony Brissonnet, EVPN : expérimentation et cas d'usages, JRES Montpellier, https://2017.jres.org/fr/presentation?id=65

[4] MLAG, https://en.wikipedia.org/wiki/Multi-chassis_link_aggregation_group

[5] JRES 2019 : Le réseau datacenter de l'UNISTRA

https://conf-ng.jres.org/2019/document_revision_5662.html?download

[6] VRF lite : instance de routage virtuelle limitée à un seul équipement réseau

https://fr.wikipedia.org/wiki/Virtual_routing_and_forwarding#Implémentation_simple