Le réseau Datacenter Unistra

Oumar Niane
Direction du Numérique
Université de Strasbourg
14 rue rené descartes
67000 Strasbourg

Christophe Palanché
Direction du Numérique
Université de Strasbourg
14 rue rené descartes
67000 Strasbourg

Fabrice Peraud
Direction du Numérique
Université de Strasbourg
14 rue rené descartes
67000 Strasbourg

Alain Zamboni
Direction du Numérique
Université de Strasbourg
14 rue rené descartes
67000 Strasbourg

Résumé

L'Université de Strasbourg (Unistra) vient d'achever la construction d'un Datacenter, d'une surface de 450 m2, répartis en quatre salles serveurs de 27 baies chacune ainsi que d'une salle opérateur de 12 baies. Pour l'urbanisation réseau de celui-ci, la Direction du Numérique (DNum) a fait le choix de partir sur un réseau indépendant de l'existant en s'appuyant sur des technologies dédiées aux datacenters.

Ce réseau est ainsi basé sur une architecture de type Spine/Leaf et fournit une séparation underlay/overlay en utilisant des protocoles eVPN/VxLAN. A l'issue d'un appel d'offres, ce modèle a été mis en œuvre sur des équipements de marque Arista. Grâce au double attachement de tous ces éléments, ce réseau permet de répondre aux contraintes de haute-disponibilité. Il permet aussi d'interconnecter le site de Plan de

Reprise d'Activité (PRA) de l'université, localisé à l'heure actuelle à Strasbourg. Les choix d'architecture effectués permettront au besoin de le délocaliser dans un autre Datacenter de l'ESR. La création de ce nouveau réseau a également été l'occasion de mettre en place des outils d'administration et de supervision à l'état de l'art.

Le poster présente les éléments composants le réseau du Datacenter, à savoir :

- *l'architecture de la Fabric ;*
- *les technologies et le matériel retenus ;*
- *les outils utilisés pour le déploiement et l'exploitation : Zero Touch Provisionning (ZTP), Api de configuration, télémétrie, sauvegarde des configurations ;*
- *des exemples de cas d'usages : prolongation d'un réseau de composante dans une baie de location, interconnexion d'un firewall redondant avec annonces de ses routes en BGP.*

Mots clés

réseau, Datacenter, eVPN/VxLAN, Fabric

1 Introduction

Pour l'urbanisation réseau du Datacenter de l'Université de Strasbourg, la Direction du Numérique (DNum) a fait le choix de partir sur un réseau indépendant de l'existant en s'appuyant sur des technologies dédiées aux Datacenters.

L'architecture répond à des exigences de modularité pour connecter les 4 salles serveurs du Datacenter composées, chacune, de 27 baies de 50U et la salle opérateur composée de 12 baies. Elle permet également l'intégration d'un site de Plan de Reprise d'Activité (PRA), l'interconnexion avec un autre Datacenter et l'hébergement d'équipements dans les baies de colocation.

2 Le *Réseau Datacenter*

2.1 Besoins et contraintes

Afin de garantir un niveau de service élevé, la DNum a souhaité proposer une redondance sur tous les éléments du réseau avec des débits aux standards actuels des Datacenters.

Ainsi, les débits ciblés à la conception étaient de :

- 100 Gb/s pour les liens de cœur de réseau ;
- 10/25 Gb/s pour les liens vers les serveurs ;
- 10/25/40/50/100 Gb/s pour les liens vers les armoires colocation.

Pour la redondance, le réseau devait permettre :

- le double-attachement de tous les serveurs à deux équipements différents ;
- la redondance de tous les liens et équipements du cœur de réseau ;
- le double raccordement au réseau métropolitain *Osiris* ;
- le double raccordement au site distant de PRA.

De plus, dans l'idée de favoriser la mutualisation des ressources, nous souhaitions disposer d'une architecture permettant à terme l'interconnexion à un Datacenter d'un autre établissement de la communauté enseignement supérieur/recherche pour remplacer notre site PRA actuel.

Enfin, nous souhaitions profiter de l'opportunité de déployer un réseau pour mettre en place des outils d'administration et de supervision à l'état de l'art.

2.2 Architecture, technologie et matériel retenus

Après avoir étudié les offres Datacenter des constructeurs, une architecture de type Spine/Leaf avec la technologie *eVPN/VxLAN* est apparue comme étant la plus répandue et adaptée à nos besoins.

e*VPN/VxLAN* est basé sur un modèle séparant plan de données et plan de contrôle :

- le plan de donnée utilise *VxLAN* qui est un mécanisme d'encapsulation des trames Ethernet sur UDP (*overlay*), permettant leur transport sur un réseau de niveau 3 (*underlay*) ;
- le plan de contrôle est réalisé par *eVPN*, qui se base sur *BGP* (*Border Gateway Protocol*) pour annoncer notamment des routes de type Mac pour les services de niveau 2 ainsi que des routes de type IP pour les services de niveau 3.

Ces technologies présentent notamment les avantages suivants :

- l'architecture Spine/Leaf garantit une latence constante grâce à un nombre fixe de sauts entre chaque élément connecté ;
- il est basé sur *BGP*, protocole de routage bien connu et éprouvé capable d'assurer le passage à l'échelle ;
- *eVPN* permet la gestion des services de niveau 2 et de niveau 3 ;
- *eVPN/VxLan* permettent l'extension d'une *Fabric* sur plusieurs sites en utilisant un simple service de niveau 3 d'un opérateur tiers ;
- La configuration des services ne se fait que sur les équipements d'extrémités, contrairement aux *VLAN* classiques qui nécessitent la configuration de tous les équipements intermédiaires ;
- la séparation en deux couches underlay et overlay réduit les risques de dommage collatéraux lors des modifications de configuration d'un service précis ;
- ces protocoles sont standardisés et implémentés par plusieurs constructeurs ;
- l'*underlay* de niveau 3 permet de s'affranchir des problèmes de boucles sur la *Fabric*, et ainsi d'éviter l'usage de protocole de type S*panning Tree* qui sont problématiques lors des extensions d'infrastructures.

À l'issue d'une phase d'étude et d'un appel d'offre (cf. Le projet *Réseau Datacenter*), c'est le constructeur *Arista* qui a été retenu pour la fourniture des équipements du réseau Datacenter. La *Fabric* est ainsi composée de trois types d'équipements différents :

- Spines : chassis Arista *7308X* (8 slots pour 256 ports 100Gb maximum)
- Leaves : Arista *7050SX3-48YC8* : 48 ports 10/25Gb et 8 ports 100Gb
- Border-Leaves : *Arista 7050CX3-32C* : 32 ports 100Gb

Le schéma ci-dessous présente l'architecture retenue du réseau Datacenter et son site de PRA

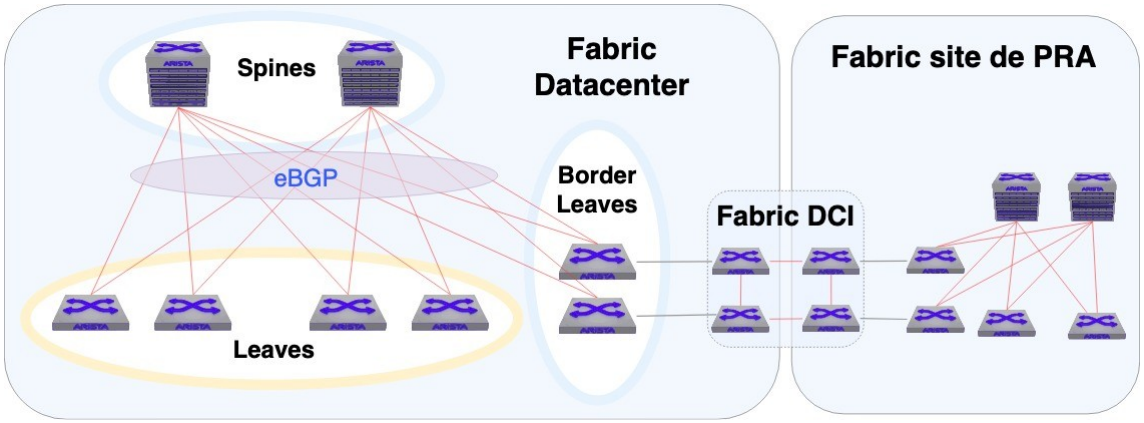

Figure 1 - Architecture du réseau Datacenter

Chacun des deux sites est équipé de deux Spines centralisant les connections des Leaves. Chaque armoire de serveurs de la Direction du Numérique est équipée de deux Leaves, chacune rattachée aux deux Spines du site. Les deux sites disposent également d'une paire de commutateurs dit "Border Leaves", qui seront les uniques points de sortie de la *Fabric*, notamment pour les cas suivants :

- les baies en location ;
- l'interconnexion *Osiris* ;
- l'interconnexion avec la *Fabric DCI* (*DataCenter Interconnection*).

L'interconnexion des sites DC et de PRA se fera par l'intermédiaire d'une *Fabric DCI*. Celle-ci est composée de deux équipements sur chacun d'eux. Elle peut être facilement étendue par le simple ajout d'équipements sur des nouveaux sites. L'extension des services vers ou depuis cette *Fabric* se fera en *802.1Q* pour les services de niveau 2 et par l'intermédiaire de réseaux d'interconnexion pour les services de niveau 3. Dans ce dernier cas, on utilisera *BGP* pour échanger les routes des services entre les différentes *Fabric*.

2.3 Outils d'administration et d'exploitation

L'administration du réseau Datacenter met en œuvre plusieurs outils. Le schéma suivant présente une vue d'ensemble de ces outils.

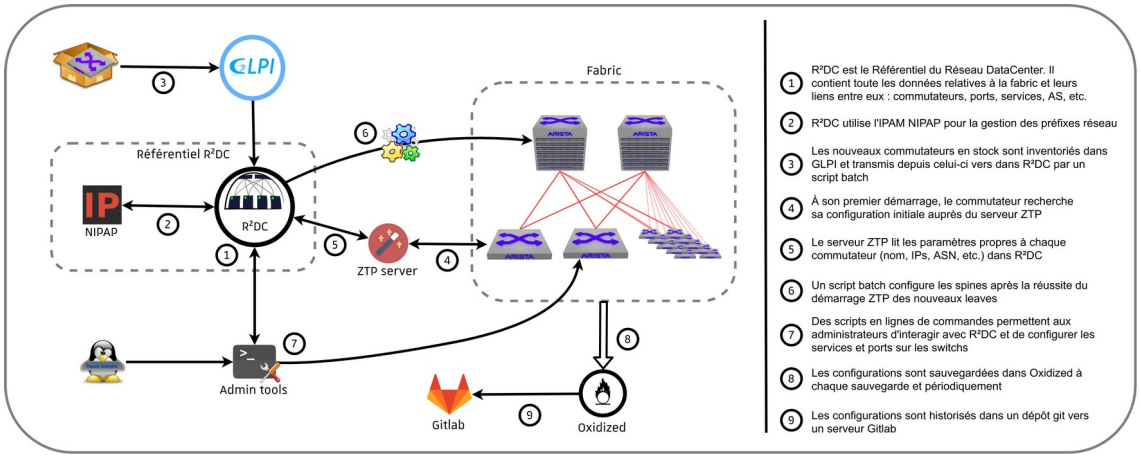

Figure 2 - Outils d'exploitation

2.3.1 Référentiel Réseau DataCenter (*R²DC*)

Pour gérer et exploiter le réseau DC, nous avons choisi de mettre en œuvre un référentiel dédié qui expose à travers une API *REST* :

- la liste des équipements réseau de la *Fabric* ;
- l'ensemble des interfaces de ces équipements ;
- les préfixes réseaux attachés aux équipements ;
- les ASN attribués aux équipements ;
- les services transportés par la *Fabric*.

R²DC est composé de plusieurs briques logicielles :

- les données sont stockées dans une base *PosgreSQL*[1] ;
- l'API, dont les spécifications sont décrites en *OpenAPI3*[2], repose sur le framework web Python *Flask*[3]. Les routes *Flask* sont automatiquement générées à partir de la spécification grâce à l'utilisation du framework *Connexion*[4], édité en open-source par la société *Zalando*. L'accès à la base de données *PostgreSQL* est effectué via l'*ORM* (*Object-Relational Mapper*) *SQLAlchemy*[5] ;
- la gestion des préfixes est déléguée à l'*IPAM* (*IP Address Management*) *NIPAP*[6]. Ses données sont accédées par les clients uniquement via *R²DC* pour garantir la cohérence entre les deux bases.

À l'heure actuelle, *R²DC* interagit avec :

- L'application d'inventaire *GLPI*, qui y injecte les équipements via un script de synchronisation

1. https://www.postgresql.org/
2. https://swagger.io/specification/
3. https://github.com/pallets/flask
4. https://github.com/zalando/connexion
5. https://www.sqlalchemy.org/
6. http://spritelink.github.io/NIPAP/

- le serveur *ZTP (Zero Touch Provisionning*) lors du processus de déploiement initial d'un équipement pour obtenir les données de configurations propres à l'équipement (noms, adresses IP, etc.) ;
- les outils de configuration des commutateurs pour y recenser les services déployés

2.3.2 Automatisation du déploiement initial

La configuration initiale des équipements de type Leaf est automatisée avec la fonctionnalité *Zero Touch Provisionning* (*ZTP*). Pour la partie serveur, nous avons déployé le serveur *ztpserver* communautaire mis à disposition sur le dépôt Github *arista-eosplus*. Celui-ci, via un module développé en interne, interroge *R²DC* pour obtenir les ressources nécessaires à la génération dynamique de la configuration de chaque équipement :

- nom de l'équipement ;
- adresse IP de management ;
- adresses IP d'interconnexion avec les Spines;
- adresses de loopback ;
- adresses IP des différents peering *BGP*;
- numéro d'AS.

2.3.3 Provisioning des services

Les différents outils de provisioning servant à configurer les services sur les équipements sont en cours de développement au moment de la rédaction de l'article. Le langage *Go* a été retenu pour réaliser ces outils, car *Arista* met à disposition sur leur dépôt *Github*[7] une librairie nommée *goeapi* relativement complète.

Ces outils doivent permettre aux utilisateurs de créer, détruire et affecter à des ports des services L2 ou L3 en fonction des droits qui sont déclarés dans *R²DC*.

Ces outils sont actuellement sous la forme d'outils en ligne de commande, mais nous envisageons à terme de proposer une interface web.

2.3.4 Sauvegarde et historisation des configurations

Les configurations des équipements sont sauvegardées par un outil de sauvegarde de configuration d'équipements réseau : *Oxidized*[8].

La sauvegarde est effectuée de deux manières différences :

7. https://github.com/aristanetworks/goeapi
8. https://github.com/ytti/oxidized

- toutes les 24H, le serveur *Oxidized* interroge tous les équipements déclarés dans sa base ;
- à chaque modification de la *startup-config*, l'équipement concerné notifie *Oxidized* pour que ce dernier déclenche une sauvegarde. Ceci est possible grâce au mécanisme d'*event-handler* proposé par les commutateurs *Arista*.

Oxidized nous permet d'historiser les configurations d'équipement avec git soit localement, soit vers un serveur externe. Nous avons choisi d'installer un serveur *Gitlab*[9] externe pour bénéficier de fonctionnalités avancées, et pour mutualiser cette historisation des configurations avec d'autres équipements non gérés par *Oxidized*.

2.4 Cas d'usages

Nous avons prévu plusieurs scénarios de raccordement au réseau Datacenter. Nous vous présentons ici les scénarios les plus courants.

2.4.1 Raccordement des serveurs de la DNum

Chaque armoire de serveurs est équipée de deux Leaves, permettant le double-attachement actif/actif de chaque serveur en *LACP*. Sur les Leaves, le double attachement est géré en utilisant des *ESI* (*Ethernet Segment Identifier*). Cette solution intégrée à *eVPN* permet de gérer les multi-homing en évitant la mise en œuvre d'autres protocoles sur les équipements tel que *MLAG* (*Multi-chassis Layer Agregation*), qui nécessite lui l'utilisation de liens dédiés entre les paires d'équipements.

De manière générale, nos serveurs sont rattachés à des services de niveau 2. Ceux-ci peuvent être routés soit sur un équipement externe à la *Fabric* (ex: firewall), soit sur la *Fabric* elle-même en *anycast-gateway* (routage au plus proche).

9. https://gitlab.com/

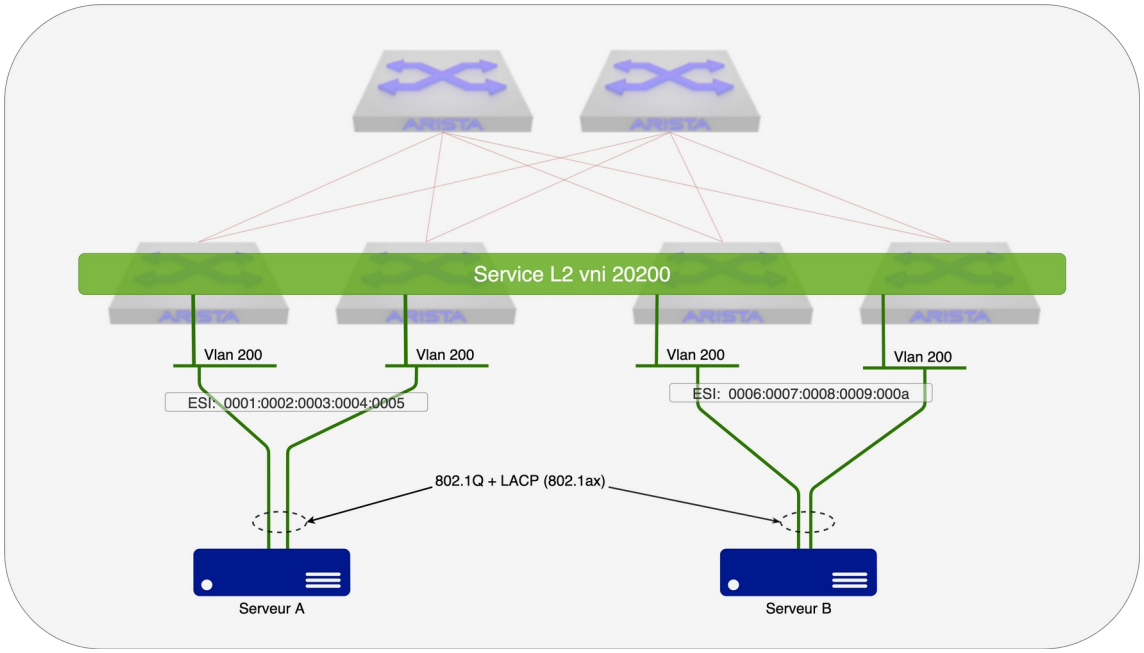

Figure 3 - Raccordement des serveurs

2.4.2 Interconnexion d'un firewall en haute-disponibilité

Chacun des firewalls sera connecté grâce à un réseau d'interconnexion unique se trouvant dans le même service de niveau 3 sur la *Fabric*. Les réseaux clients seront directement rattachés aux firewalls à travers des services de niveau 2 et le firewall master sera désigné grâce au protocole *CARP*. Le firewall master annonce ses différents réseaux à la *Fabric* en *eBGP*.

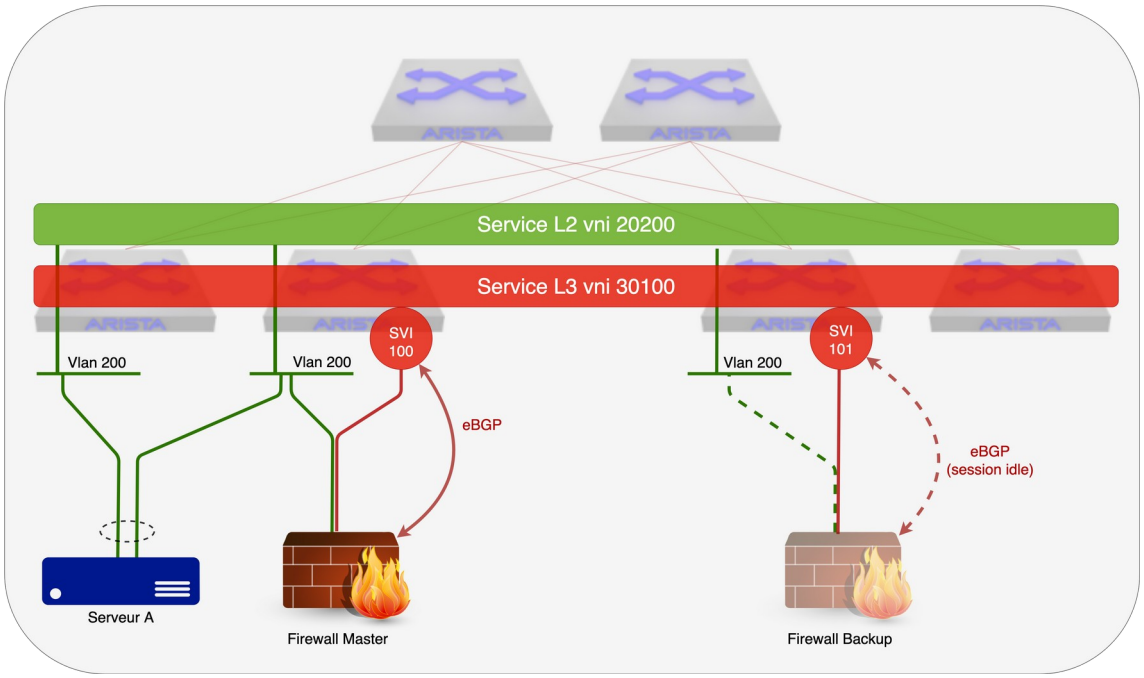

Figure 4 - Interconnexion d'un firewall en haute-disponibilité

2.4.3 Interconnexion d'une baie de colocation

À l'heure actuelle, notre offre de colocation inclut la fourniture d'une connexion redondante au réseau en *802.1Q*. Dans leurs armoires, les clients doivent gérer la distribution du réseau sur leurs propres équipements.

Concernant les réseaux utilisés dans les armoires louées, les clients peuvent disposer de réseaux transverses à leurs différentes armoires dans le Datacenter. Ceux-ci peuvent aussi être prolongés en niveau 2 jusqu'au point d'entrée d'un bâtiment raccordé au réseau métropolitain *Osiris*.

3 Le projet "Réseau Datacenter"

La mise en place du Réseau Datacenter a pris la forme d'un sous-projet du projet global *Datacenter Unistra*. Il avait pour objectif la fourniture et l'installation du réseau dans le nouveau Datacenter de l'Unistra et de son site de PRA.

Le choix ayant été fait d'urbaniser deux salles sur quatre à l'ouverture du Datacenter, le projet consistait donc à connecter au réseau 29 baies dans le Datacenter et 6 sur le site de PRA.

Ce chapitre présente les différentes phases du projet.

3.1 Planning

Le projet s'est étendu sur une période d'environ deux ans et demi. Il peut être découpé en trois grandes parties :

- l'étude et le choix des solutions techniques ;
- la phase d'appel d'offres ;
- la préparation et le déploiement du matériel.

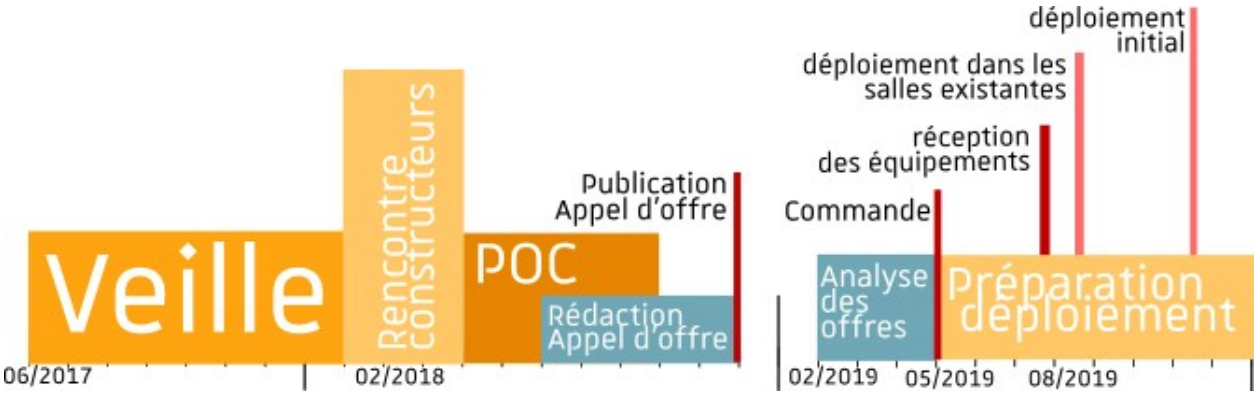

Figure 5 - Planning

3.1.1 Sourcing

Cette première étape du projet nous a permis de définir l'architecture générale et les technologies que nous allions mettre en œuvre pour le réseau du Datacenter.

Dès 2017, nous avons entrepris une démarche de veille sur les architectures et technologies recommandées pour les réseaux de Datacenter. À partir de février 2018, nous avons rencontré les principaux constructeurs de solutions réseaux Datacenter afin de :

- découvrir leurs gammes d'équipements ;
- confirmer les résultats de nos recherches ;
- affiner nos choix d'architecture :
- sélectionner quelques constructeurs pour effectuer des POC (Proof Of Concept).

À l'issue du sourcing, nous avons pu confirmer que l'architecture Spine/Leaf et la technologie *eVPN/VxLAN* répondaient à nos besoins et étaient les plus mises en avant par les constructeurs sur ce segment de marché.

3.1.2 Proof Of Concept

À l'issue du sourcing, de mai 2018 à octobre 2018, quatre constructeurs nous ont prêté du matériel pour réaliser une maquette représentant l'architecture retenue pour le réseau Datacenter.

Celle-ci nous a permis de :

- prendre en main la technologie retenue et d'en approfondir nos connaissances ;
- tester plusieurs cas d'usages représentatifs de notre future offre de service ;
- identifier les fonctionnalités importantes à demander dans notre cahier des charges d'appel d'offres.

3.1.3 Appel d'offres

La période d'appel d'offres s'est étendue de juillet 2018 à avril 2019. Sa préparation formelle a chevauché la fin des POC, de juillet à novembre 2018. Le marché a été publié en décembre 2018 pour une attribution le 5 avril 2019.

L'évaluation financière des offres s'appuyait sur un devis quantitatif et estimatif représentant l'équipement nécessaire à l'ouverture du Datacenter.

Pour toutes les offres, chaque type d'équipement a été jugé par rapport à une liste de fonctionnalités exigées. Chaque solution a été évaluée dans son ensemble suivant une série de tests effectués sur du matériel prêté par le soumissionnaire. Un cahier de tests inclus dans le dossier d'appel d'offres recensait l'architecture de la maquette et les tests effectués. Pour garantir l'équité vis-à-vis des constructeurs non familiers de l'équipe Unistra, il était demandé aux soumissionnaires de fournir cette maquette fonctionnelle et des modèles de configuration pour chaque test. Ces mesures ont ainsi permis

l’utilisation de configurations optimales fournies par le soumissionnaire pour chacun des tests.

Dès l’attribution du marché, la commande initiale a été finalisée et envoyée au candidat retenu.

3.1.4 Préparation et déploiement

En attendant la livraison du matériel, l’équipe projet a pu commencer la préparation des configurations des équipements pour la première vague de déploiement.

Pour profiter des débits de 10/25 Gb/s proposés par le nouveau réseau Datacenter, les cartes réseaux de nombreux serveurs devaient être remplacées ou reconfigurées. Afin de décorréler ces opérations et le déplacement physique des serveurs dans le Datacenter, la Direction du Numérique a décidé de déployer par anticipation le réseau Datacenter dans les salles serveurs historiques.

De mai à août 2019, l’équipe projet a donc œuvré pour préparer une *Fabric* étendue à destination des anciennes salles serveurs. Pour simplifier les opérations de déménagement des serveurs cette *Fabric* sera étendue au Datacenter.

En parallèle, l'équipe a commencé à travailler sur les différents outils d'administration : le référentiel *R²DC*, les outils de déploiement et les outils de sauvegarde et d’historisation des configurations.

Depuis septembre 2019, elle prépare le déploiement des équipements destinés aux salles du Datacenter, qui devraient être en production au moment des JRES.

4 Retour d'expérience

À la date de rédaction de cet article, la *Fabric* n’a pas encore été déployée dans le Datacenter suite à des retards dans la livraison du bâtiment. Notre retour d’expérience porte donc sur la *Fabric* déployée dans les salles serveurs historiques en août 2019. Le dimensionnement actuel de ces salles ne nous permet donc pas d’avoir un regard sur les gains potentiels de la *Fabric* sur le passage à l’échelle.

Cependant, nous avons déjà pu relever un net gain de performance grâce à l’augmentation du débit (passage de 1 Gb/s actif/passif vers 10/25 Gb/s actif/actif) et la baisse des latences. Cette amélioration bénéficie particulièrement aux infrastructures de stockage *CEPH*.

À ce stade, nous n’avons eu aucun incident à déplorer sur la *Fabric*.

5 Travaux en cours et perspectives d’évolutions

5.1 Fabric DCI et site de PRA

Au moment des JRES, seul le site principal du Datacenter aura été déployé. Notre prochain objectif en matière d'infrastructure est maintenant de déployer le site de PRA et par conséquent la *Fabric DCI*. Ces opérations devront être terminées avant d'investir ce site en mars 2020.

5.2 Évolutions des outils d'exploitation

Une feuille de route a été établie afin de prioriser les fonctionnalités de nos outils d'exploitation. En effet, ils ne couvrent actuellement que les fonctionnalités qui étaient nécessaires au déploiement initial du réseau DC et sa configuration pour accueillir les serveurs existants de l'Unistra :

- dans *R²DC*, les routes nécessaires à l'auto-configuration *ZTP* ;
- import des commutateurs dans *R²DC* depuis *GLPI* ;
- script en ligne de commande pour créer des services de niveau 2 et affecter des *VLAN* aux ports.

Par la suite, nous prévoyons plusieurs évolutions afin d'étoffer l'offre de service de réseau DC et de rendre l'architecture logicielle plus scalable.

5.2.1 Configuration des commutateurs

Les scripts actuels, écrits en *GO*, transmettent directement des paramètres de configuration aux commutateurs via leur API web. Dans le futur, nous envisageons plusieurs évolutions à ce fonctionnement :

- mise en place d'une API centralisant les demandes de configurations de switchs. Cette API proposerait les opérations courantes d'exploitation de la *Fabric*. Elle interagirait avec *R²DC*, pour, entre autres, consulter et mettre à jour les informations du référentiel réseau Datacenter ou encore vérifier les autorisations du client ;
- mise en place d'une file de message, transmettant au commutateur les configurations ;
- déploiement d'un client de configuration sur chaque commutateur. Ce client serait chargé d'appliquer les modifications de configurations lues sur la file de message.

Ces évolutions permettraient une gestion asynchrone des configurations, qui s'avérerait plus efficace dans l'optique de proposer une interface web de paramétrage de la *Fabric*.

5.2.2 R²DC

En vue de déléguer l'utilisation de nos outils de configurations des services sur la *Fabric* à nos clients, la gestion des autorisations et des droits sera implémentée dans *R²DC*. Cette évolution vise à restreindre l'accès via l'API aux ressources et méthodes proposées par R²DC selon l'identité du client.

Enfin, nous envisageons la possibilité d'intégrer directement les fonctionnalités d'*IPAM* dans *R²DC* et de ne plus utiliser *NIPAP*. En effet, si *NIPAP* nous a permis de gagner du temps dans la création du *R²DC*, nous devons maintenant en contrepartie garantir une cohérence entre les deux bases (liens entre objets ou informations dupliquées). Ceci complexifie certaines opérations de maintenances (correction de bugs, migrations et restauration de données, etc.).

5.2.3 Interface web

Pour aller plus loin dans la délégation des opérations de configuration, nous souhaitons mettre en place une application web offrant une vue sur les services et les équipements de la *Fabric*. Elle permettrait également de réaliser certaines opérations de configurations, comme la création des services ou leur affectation. Elle s'appuierait sur la gestion de droits implémentée dans R²*DC* pour limiter les accès aux ressources et aux opérations.

5.3 Télémétrie (supervision/métrologie) avec InfluxDB, ELK, Grafana

Pour le moment, nous n'avons pas encore mis en place la télémétrie sur notre infrastructure.

A l'issue des tests réalisés pendant le POC et la phase d'appel d'offres, nous pensons nous orienter vers les solutions suivantes :

- *InfluxDB* pour stocker les données métriques ;
- une stack *Elasticseach/Logstash/Kibana (ELK)* pour le stockage des logs et des événements ;
- *Grafana* pour la visualisation de ces éléments.