MASTER EN SCIENCES INFORMATIQUES

Placement des SFC
recensement critique de la littérature et perspectives

DE ROP, Christophe

Award date:
2023

Awarding institution:
Universite de Namur

UNIVERSITÉ DE NAMUR
Faculté d'informatique
Année académique 2022-2023

Placement des SFC : recensement critique de la littérature et perspectives

Christophe DE ROP

..... (Signature pour approbation du dépôt - REE art. 40)

Promoteur : Laurent SCHUMACHER

Mémoire présenté en vue de l'obtention du grade de Master 60 en Sciences Informatiques

Faculté d'Informatique – Université de Namur

RUE GRANDGAGNAGE, 21 ● B-5000 NAMUR(BELGIUM)

Chapitre 1

Remerciements

Ce travail est la dernière étape de mon parcours académique. Quand je regarde les difficultés rencontrées, je suis particulièrement fier d'être arrivé à cette ultime étape que représente la rédaction de ce mémoire.

Je tiens à exprimer toute ma reconnaissance à mon promoteur, Monsieur Laurent SCHUMACHER, Professeur et Vice-Recteur de l'Université de Namur, pour sa patience, sa disponibilité et sa bienveillance. Il m'a prodigué de bons conseils et des remarques judicieuses qui m'ont permis de progresser dans ce travail.

Je tiens également à remercier mon épouse, Stéphanie MOISSE qui m'a soutenu et encouragé en toutes circonstances. Un tout grand merci à mes deux enfants, Alexandre et Marie, pour leur soutien pendant les moments difficiles, pour leur patience durant mes absences. J'espère qu'ils pourront apprendre de cette expérience et par l'exemple ce que les mots des parents n'arrivent pas toujours à exprimer.

Je remercie Madame Sandrine DE ROP, pour sa relecture et sa correction de mon travail.

Je remercie tous mes collègues qui ont dû subir mes absences pour cause d'examens et autres sessions d'études. Je suis conscient de l'effort qu'ils ont consenti pour arranger leurs horaires et me permettre de poursuivre ces années académiques dans les meilleures conditions possibles. Un merci particulier à Monsieur Laurent GILLARD pour sa flexibilité et à Monsieur Michael FALLAY pour son aide et ses conseils dans certains travaux.

Enfin, je remercie toute ma famille et ma belle-famille pour leurs encouragements.

Chapitre 2

Résumé

La multiplication des applications mobiles, la démocratisation de l'utilisation d'Internet, les services offerts sur ce réseau, l'augmentation d'objets connectés a été permise par l'adaptation des réseaux informatiques et leurs capacités à transmettre les données de manière satisfaisante. Un service doit franchir une série d'étapes situées entre l'utilisateur et le gestionnaire du service dans des conditions définies de performance, de disponibilité et de sécurité. Cette suite forme une *Service Function Chain*. Routeurs, switches, firewalls, proxies et autres outils de sécurité ou d'amélioration des performances étaient jadis des boîtiers intermédiaires physiques offrant des solutions propriétaires. Si certains demeurent, d'autres sont aujourd'hui virtualisés dans des environnements dédiés spécifiquement à la gestion du cycle de vie des machines virtuelles, le *Network Function Virtualisation*.

Avec la dématérialisation des boîtiers intermédiaires, des algorithmes d'optimisation du placement de ces boîtiers virtuels sont apparus pour permettre la meilleure orchestration possible. Le placement a été initialement défini via une fonction de coût dans le cadre d'une optimisation linéaire en nombres entiers. Ensuite, le passage à l'échelle a orienté la recherche vers des algorithmes heuristiques, plus efficaces pour rendre une réponse sous-optimale à ce problème de placement et pouvant s'orienter vers certains aspects spécifiques. Ce travail rappelle les différents aspects du placement des machines virtuelles évoqués dans la littérature. A travers ces articles, on notera également les évolutions de l'environnement dans lequel s'insèrent ces fonctions virtualisées : *cloud-computing*, *edge-computing*, 5G et 6G, augmentation du trafic qui impactent la prise de décisions de ces algorithmes. L'intelligence artificielle propose des méthodes en prenant en compte des contraintes telles que les spécifications des réseaux du futur (5G, 6G), la prédiction du comportement des réseaux et la prise de décisions automatisée sur l'apprentissage.

A la lecture des différents articles utilisés dans ce document, des problèmes sont également soulevés sur l'emploi ou la faisabilité de certaines méthodes. Sur base d'autres auteurs, des réponses critiques sont formulées sur ces questionnements et des propositions sont également évoquées pour avancer sur certains points.

Chapitre 3

Index

Table des matières

Table des figures

Liste des tableaux

Chapitre 4

Introduction

Aujourd'hui, nous pouvons constater que la numérisation des services est largement installée dans notre société. Nous avons acquis le réflexe de rechercher nos informations via un moteur de recherche, de contacter entreprises et administrations par mail, par chatbox, de prendre directement nos rendez-vous en accédant à des agendas en ligne. Nos prescriptions médicales peuvent être envoyées directement sur un serveur consultable par notre pharmacien ou le patient. Nos voitures ont des fonctionnalités d'orientation, de gestion de trafic, de sécurité, de confort d'utilisation tout comme les retards de nos trains et nos bus nous sont signalés. La position de notre transport Uber est géolocalisée sur l'application mobile. Nous pourrions aussi rappeler les services de paiement depuis nos GSM, des services d'authentification comme Itsme, de streaming à la demande, de jeux, le télétravail avec des collègues géographiquement éloignés, la remontée automatique de données de nos compteurs électriques ou notre déclaration d'impôts pré-remplie. La pandémie de COVID a montré l'utilité des nouvelles technologies comme la visioconférence et les outils collaboratifs. Malheureusement, l'invasion de l'Ukraine montre aussi que les moyens informatiques sont largement exploités, que ce soit dans la guerre hybride avec le hacking des infrastructures adverses, l'utilisation des réseaux de satellites pour la communication (par exemple Starlink) ou pour le renseignement mais aussi le discours mettant en avant de nouveaux concepts comme la fusion de données venant de multiples sources pour assister à la prise de décision sur le terrain [17] ou celui de drones volants en essaim assistant des plateformes avec des pilotes humains [18] quand ce ne sont pas les drones eux-même qui définissent la meilleure cible. Les utilisateurs de ces services attendent de ces derniers fiabilité, sécurité et disponibilité alors que les opérateurs de ces services recherchent une rentabilité au terme du développement, de la mise en production et de la maintenance. Les promesses de performance de la 5G [19] ouvrent aussi la voie à la réalité augmentée "en ligne" pour les touristes visitant une ville ou pour un architecte présentant son projet sur site, la médecine à distance avec les consultations en ligne ou l'opération chirurgicale télécommandée depuis un autre endroit, ...

L'utilisation de ces applications nécessite des infrastructures qui doivent supporter le trafic nécessaire à leur bon fonctionnement, les différentes demandes qualitatives des clients, que cela soit en terme de performance, de sécurité, de disponibilité, de débit. Les organismes en charge de ces infrastructures ont à leur disposition un ensemble de systèmes divers et variés pour organiser ces trafics. Parmi les plus connus, citons de manière non exhaustive les routeurs, les switches, les firewalls, les proxys, les NATs, les antivirus, les IDS et les IPS. Lorsqu'une connexion s'établit entre un client et un serveur, l'ensemble des systèmes traversés par le trafic résultant de cette connexion forme une chaîne. Cette chaîne, appelée Service Functions Chain (SFC), est une suite ordonnée de systèmes informatiques traversés par le trafic. Ce trafic passe par des machines physiques propriétaires dont l'optimisation du paramétrage nécessite souvent une expertise propre. Ceci se traduit généralement par une utilisation commune de ces matériels pour l'ensemble du trafic. L'ordre des appareils traversés par ces flux de données relève de l'expérience des gestionnaires des infrastructures traversées. En simplifiant le propos, tant que les fonctions sont liées à un matériel physique particulier, la maintenance demande du personnel bien formé, les modifications de fonctionnement peuvent amener à des effets indésirables, les ajouts ou suppressions de fonctions réclament une préparation lourde. Cela prend du temps pour un coût élevé, le tout dans une infrastructure particularisée par l'opérateur.

Ces obstacles ont pu être levés par l'apparition de la virtualisation et la volonté de transformer toutes ces machines en équivalents virtualisées, les Virtual Network Functions (VNF). L'idée maîtresse est de créer, maintenir en bon état de fonctionnement, modifier, voire arrêter dans une infrastructure de virtualisation un ensemble de fonctions virtuelles, de les connecter les unes aux autres dans un ordre déterminé sur une simple requête et de permettre une orchestration centralisée de cet ensemble. Cela aboutirait à une gestion de plus haut niveau des SFC et permettrait une meilleure gestion des ressources informatiques et énergétiques mobilisées de manière sous-jacente par le biais de la virtualisation. La norme RFC 6775 nous donne un modèle généraliste de la SFC. Une architecture logique de ce type d'environnement est proposée par l'European Telecommunications Standards Institute (ETSI). Une architecture d'interconnexion entre différents centres de données distants (Network Function Interconnect (NFIX)) est proposée, elle, par l'Internet Engineering Task

Force (IETF).

L'implémentation de ces concepts pour former un environnement propice au déploiement des SFC est difficile. Parmi les difficultés rencontrées, il y a la gestion de l'arrivée imprévisible d'une requête nécessitant une SFC particulière et de la vérification de l'ensemble de la structure pour permettre la mise en place de cette SFC sans affecter les éventuelles autres SFC déjà présentes. Par la suite, il y a la nécessaire balance entre le maintien de la disponibilité des éléments constitutifs d'une SFC et la désactivation de la SFC lorsque cette dernière n'est plus utilisée. Ce retrait se justifie pour des raisons d'économie de moyens. Au même titre que la gestion des VNF, il y a aussi la gestion de la disponibilité des ressources matérielles de l'environnement supportant la virtualisation pour créer les VNF nécessaires. Un des axes de la recherche porte sur le partage de VNF communes à plusieurs SFC moyennant une modification de configuration. Dans un scénario de clients mobiles, ces derniers peuvent être amenés à changer de réseau. La promesse du maintien des services en cours de fonctionnement peut entraîner des modifications des SFC en fonction des acteurs gérant tout ou partie du réseau supportant ces services. Dans l'optique de rencontrer l'efficacité attendue des services, de gérer un nombre grandissant de services tout en étant rentable pour les différents acteurs des différents réseaux traversés par les flux nécessaires à ces services, le placement optimal des VNF est un enjeu important. Il doit répondre aux questions de performance, de sécurité, de disponibilité attendues de la SFC ainsi que celles relatives au coût pour l'opérateur.

La recherche documentaire reprend à travers les articles trouvés plusieurs familles d'algorithmes allant des algorithmes pour résoudre des problèmes d'optimisation linéaire en nombres entiers jusqu'à l'utilisation de l'intelligence artificielle pour l'optimisation du placement des VNF, en ligne et prédictive. Les auteurs de cette littérature montrent parfois certaines limites de la solution proposée. Ces textes couvrent de multiples cas de recherche et de domaines comme la 5G, l'emploi de drones, un centre de données, l'océanographie. Il faut nécessairement analyser les résultats et les conclusions de ces études pour s'interroger sur l'utilité de certaines méthodes.

Pour éviter la confusion entre une version française traduite et son pendant original en anglais, les acronymes anglais seront systématiquement utilisés, même dans les citations traduites. Une liste des acronymes est disponible au chapitre 9.

Chapitre 5

État de l'art

5.1 Service Function Chain (SFC)

"Une chaîne de fonctions de service (SFC) est une séquence ordonnée de fonctions de réseau (NF)." [20], "...les flux de trafic doivent passer par plusieurs étapes de middleboxes dans un ordre particulier..." [10], "...le trafic utilisateur entrant doit souvent passer par un sous-ensemble de fonctions de réseau dans un ordre spécifique..." [21], "...un ensemble de fonctions virtuelles de réseau à exécuter selon un ordre donné..." [22], "Pour un service donné, la vue abstraite des fonctions de service requises et l'ordre dans lequel elles doivent être appliquées" [2], "Ces VNF, lorsqu'elles sont enchaînées dans un ordre de traitement strict,..." [23] et d'autres phrases présentes dans d'autres articles donnent une définition de la SFC.

La SFC est un concept bien connu pour les gestionnaires d'infrastructures. Initialement, elle représente une suite de "middleboxes" physiques utilisées par les gestionnaires d'infrastructures pour offrir des services de sécurité, de support ou d'amélioration du trafic réseau à une requête tout en rendant la meilleure qualité de service (QoS) possible aux utilisateurs de cette chaîne. La littérature cite régulièrement les firewalls, NAT et autres load balancers, IDS, IPS, ... Or, ces middleboxes physiques sont des systèmes propriétaires, avec leurs caractéristiques physiques, leurs technologies matérielles et logicielles propres. Il est donc logique que la littérature souligne les difficultés rencontrées dans l'implémentation, la gestion, la maintenance, l'évolution, l'inter-opérabilité et les coûts générés par l'ensemble de ces appareils formant cette infrastructure physique [10, 24–30].

Une conceptualisation plus formelle d'une SFC a été proposée en 2015 via la RFC 7665 [1]. Elle formule de manière abstraite et générale la SFC, son fonctionnement, ses limites et les points à explorer. Si on s'en tient à la définition de la SFC proposée dans ce document, celle-ci est *"un ensemble ordonné de fonctions de service abstraites et de contraintes d'ordre qui doivent être appliquées aux paquets et/ou aux trames et/ou aux flux sélectionnés à la suite de la classification."* Sur base de cette définition, l'image d'un graphe orienté, où les noeuds sont les fonctions du réseau et les arêtes sont les liaisons entre ces fonctions, nous vient rapidement à l'esprit. Nous retrouvons les expressions tournant autour de l'idée de *"séquence ordonnée de fonctions de réseaux"* déjà évoquéé précédemment, notamment par [20]. Halpern et Pignataro [1] vont plus loin dans la définition de la SFC en ajoutant la notion de classification. Celle-ci est définie dans leur article comme étant *"la correspondance instanciée localement des flux de trafic par rapport à la politique pour l'application ultérieure de l'ensemble requis de fonctions de service réseau. La politique peut être spécifique au client/au réseau/au service"*. Dans leur proposition d'architecture, cette classification est le point d'entrée de toute SFC. Elle reçoit les paquets entrants et, suite à un processus de classification, oriente ces paquets vers une SFC particulière. La définition d'Halpern et Pignataro montre que la SFC est un ensemble ordonné d'étapes traversées par des flux, avec une série de conditions qui caractérise une SFC et la différencie par rapport aux autres. Ces contraintes conditionneront le placement des fonctions de services dans l'infrastructure.

Le document explicite le modèle d'architecture et apporte des précisions sur les actions exécutables sur les différentes fonctions réseaux abstraites. Afin de lever une éventuelle ambiguïté, une fonction de service abstraite représente la fonction offerte, sans description de la manière dont la fonction est implémentée. Par exemple, si la fonction de service est un firewall, cela englobe tous les firewalls possibles : une middlebox physique, un logiciel tournant sur une machine standard ou une version virtualisée. Le modèle autorise le caractère unidirectionnel ou bidirectionnel d'une SFC, c'est-à-dire que le chemin pour la requête est différent de la réponse. Faire des cycles est également possible. Plusieurs politiques peuvent s'appliquer à plusieurs trafics sur une même SFC. La concaténation de différentes SFC est également valide pour autant que toutes les règles des différentes SFC soient respectées. Sont définis dans l'article [1] plusieurs services nécessaires à la prise en charge complète de la SFC, comme l'état de la fonction réseau, du lien, le contrôle des politiques, le saut suivant.

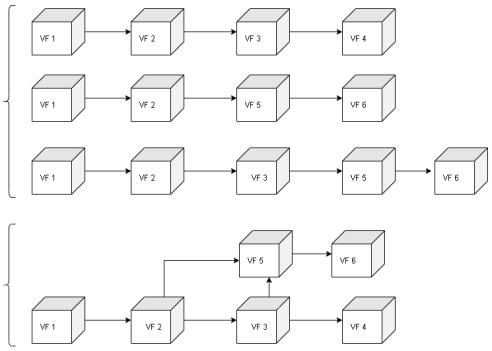

FIGURE 5.1 – Schéma représentant l'optimisation de plusieurs SFC autorisée par l'architecture et inspirée de [1]

L'enchaînement des composants d'une SFC est ordonné. Les fonctions réseaux ont des impacts sur le comportement du trafic réseau. Un enchaînement différent des composants peut produire un résultat différent de celui attendu. Certains composants ralentissent le trafic réseau, d'autres l'accélèrent, d'autres encore vont exclure une partie du trafic. Tout l'art d'une SFC efficace est le placement optimal de ces fonctions les unes par rapport aux autres pour respecter les demandes de chaînes et certaines de ces demandes peuvent nécessiter un déplacement d'un NF dans la SFC [5]. Avec des matériels réseaux physiques composant la chaîne, ce type de flexibilité est difficilement atteignable. Avec du matériel réseau virtualisé, cette flexibilité est envisageable et permet la création de nouvelles SFC plus spécifiques [11]. La virtualisation impose un environnement adapté à la création, modification, suppression de machines virtualisées et équipé d'outils de contrôle, de maintenance, centralisés. Idéalement, il faudrait que cet environnement fonctionne sur du matériel non propriétaire et aux standards de l'industrie.

5.2 Network Functions Virtualization (NFV)

A la sortie de [1], cet environnement est déjà en cours de conception. Dès 2012, soit 3 ans avant la publication de la RFC-7665, l'ETSI crée le groupe de travail sur le NFV (ETSI ISG NFV) et publie son *white paper "Network Functions Virtualisation – Introductory White Paper"* [31] où il définit les buts de la recherche menés par les futurs travaux. En effet, l'utilisation des middleboxes physiques présentait quelques inconvénients déjà constatés par les administrateurs de ces équipements : manque de flexibilité, décentralisation de la gestion, faible tolérance aux pannes matérielles, manque de souplesse pour absorber les évolutions [5, 10, 12]. Ce *White Paper* [31] traite les buts de la virtualisation des middleboxes physiques, de leur emploi dans un environnement propice et rentable, reposant sur du matériel informatique standard avec comme point de mire, la réduction de la consommation électrique et des coûts d'équipements, les économies d'échelle au passage à des middleboxes virtualisées, la facilité de gestion des services en fonction (arrêt, augmentation de capacité, optimisation, intégration de nouvelles fonctions, ...) ainsi que l'ouverture vers de plus petits acteurs. L'ETSI est une organisation supportée par plus de 900 membres dont des gros acteurs industriels ([32]). Le groupe ETSI ISG NFV comprend 79 membres et 54 participants [33]. Ceci montre qu'il y a un intérêt certain de la part des industries du secteur des télécommunications pour la mise en place de standards dans ce domaine particulier.

Les travaux de l'ETSI à travers son groupe ETSI ISG NFV ont permis de créer une architecture NFV et un cadre considéré comme une référence [2]. Cette architecture est conçue tant pour les centres de données que pour les fournisseurs de services [12]. Un nombre significatif des documents découverts lors de la recherche documentaire (voir chapitre 6.6) font directement référence aux travaux de définition, de standardisation ou de développement du logiciel OSM. [5, 6, 9–11, 13, 22, 24, 25, 27, 28, 30, 34–41]. Deux bémols sont à signaler. Le premier vient des performances des VNF comparées à leurs équivalents physiques. L'article [12] indique que les performances sont, en général, du côté des matériels physiques. Le second bémol se situe dans l'article de [9] où les auteurs signalent qu'en dépit des progrès, *"il reste encore à démontrer l'existence d'une plateforme MANO à part entière conforme à la norme ETSI, y compris les opérations d'orchestration dynamique"*. Les informations de ces deux bémols datant un peu, elles sont à vérifier.

L'architecture de l'ETSI telle que présentée à la figure 5.3 montre 3 zones et 13 blocs fonctionnels. De manière très généraliste, nous avons :

— La zone "VNF" où se trouvent les fonctions réseaux virtualisées (VNF) remplaçant les middleboxes physiques et leurs gestionnaires d'élément (EM). Chaque gestionnaire d'éléments remonte l'état de sa VNF.

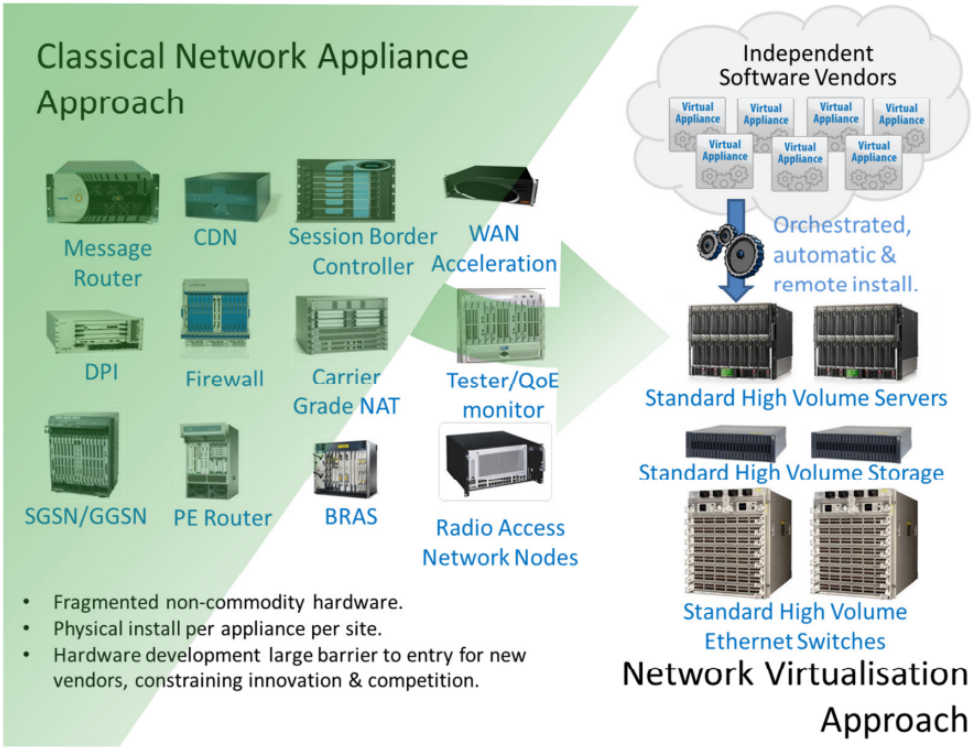

FIGURE 5.2 – Vision pour NFV - White Paper ETSI, octobre 2012

— La zone "NFV Infrastructure" ou NFVI, qui fournit à la zone "VNF" les ressources nécessaires pour fonctionner. On y retrouve les ressources matérielles, à savoir les processeurs, la quantité de mémoire et l'espace de stockage ainsi que les composants réseaux physiques, comme les switches. Également présent dans cette zone la couche de virtualisation avec son logiciel de virtualisation. C'est à travers cette couche que l'on propose les ressources virtuelles nécessaires à la création des machines virtuelles (VNF de la zone VNF). Pour se faire une idée, ce serait des serveurs lames connectés à une baie de disque de type SAN. Les serveurs ont un logiciel de virtualisation installé qui donne les outils pour créer des machines virtuelles.

— La zone "NFV Management and Orchestration" qui s'occupe de la gestion des ressources logicielles et physiques des zones VNF et NFVI. C'est à ce niveau qu'arrive une requête. On peut donner un aperçu du fonctionnement de cette zone. Elle est mise sous forme de graphe reprenant les différents composants nécessaires pour rencontrer la demande formulée par la requête. On intègre aussi les règles imposées par le bloc "OSS/BSS" et les contraintes imposées par la requête. Un algorithme va ensuite calculer sur base de l'existant, des ressources nécessaires et des ressources disponibles, l'emplacement optimal des machines virtuelles (VNF). Ce processus est appelé "placement des VNF". Si une solution est trouvée, les ordres de création, de modification, de déplacement des machines virtuelles sont envoyés et exécutés. La SFC est ainsi mise en place et fonctionnelle. Ces tâches sont prises en compte par 3 blocs fonctionnels distincts.

  — Le gestionnaire d'infrastructure virtualisée (VIM) gère le bloc NFVI. Il remonte les statuts des différents composants de ce bloc vers l'orchestrateur. Il attribue des ressources disponibles à la création ou à la modification des machines virtuelles. C'est à travers ce bloc que l'orchestrateur a connaissance du monde physique sur lequel repose la partie virtuelle. Conceptuellement, il peut y avoir plusieurs VIM pour un même orchestrateur.

  — Le gestionnaire de fonctions de réseaux virtualisés (VNF Manager) se charge de la partie logicielle embarquée sur les machines virtuelles. Les gestionnaires d'éléments (EM) transmettent les données d'état de leurs VNF au gestionnaire de VNF qui les centralise et les remonte à l'orchestrateur. Ces informations permettent la gestion du cycle de vie de la SFC par l'orchestrateur. Tout comme le VIM, plusieurs VNF managers peuvent être connecter à un orchestrateur.

  — L 'orchestrateur NFV (NFV-O ou NFV Orchestrator) centralise les informations des autres blocs, les interprète et le cas échéant, prend action afin de respecter les contraintes commerciales et opérationnelles définies dans le bloc "OSS/BSS". Les actions possibles sont l'ajout/la diminution de ressources à une VNF/SFC, l'arrêt de la

SFC, le déplacement d'une VNF, la création d'une nouvelle VNF sur une SFC existante, l'ajout d'une nouvelle SFC, la concaténation d'une SFC avec une autre SFC existante.

— Le bloc "OSS/BSS" est le centre des règles. C'est ici que l'on définit les règles commerciales, les règles opérationnelles et la logique de contrôle de la chaîne de services.

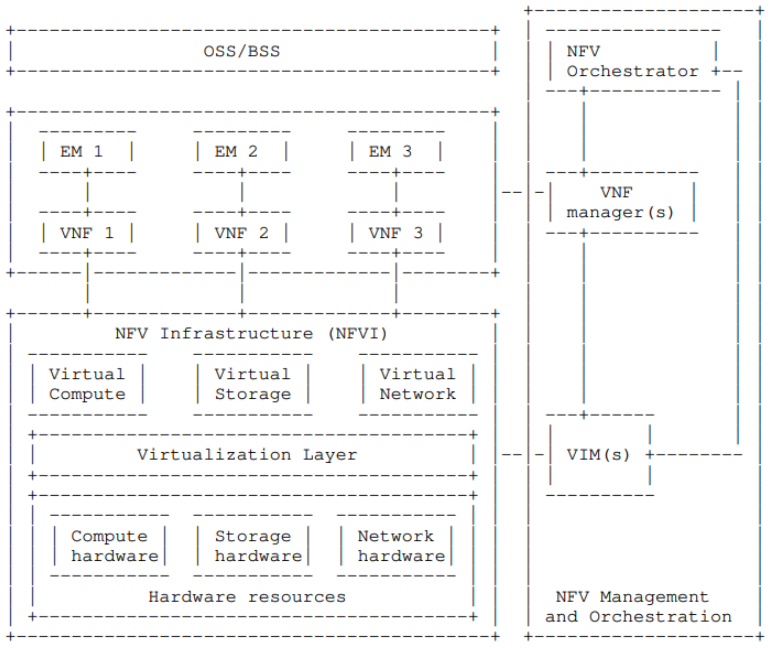

FIGURE 5.3 – NFV ETSI Architecture [2]

Suite à cette description, plusieurs points remarquables de cette architecture sont à mettre en évidence.

Le premier point est la centralisation de la prise de décision au sein de l'orchestrateur sur l'ensemble de l'architecture. En effet, il est explicitement montré sur le graphique 5.3 qu'un orchestrateur peut gérer plusieurs VNF managers et plusieurs VIM. Dit autrement, un orchestrateur peut gérer plusieurs infrastructures physiques de virtualisation dont les distances géographiques peuvent être grandes mais les considérer comme une entité logique unique. Cette évolutivité de l'architecture permet de prendre en compte des petites structures comme des grandes et de leur permettre de grandir/diminuer par la suite. Le modèle de l'époque du graphique 5.3 a légèrement évolué pour prendre en compte, dans la zone "NFV Management and Orchestration", un *"Wide Area Infrastructure Manager (WIM)"*.

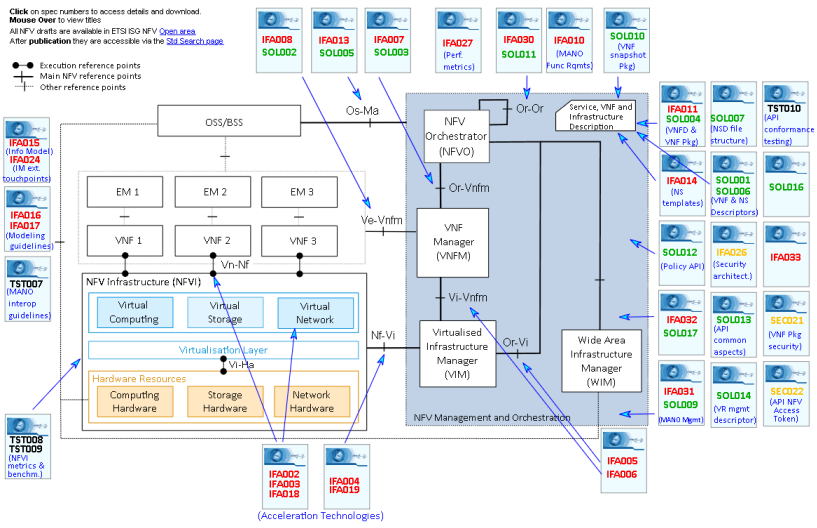

FIGURE 5.4 – Architecture NFV disponible sur le site de l'ETSI [3]

Le second point est le caractère agnostique des logiciels utilisés pour les VNF managers et VIM. L'architecture se contente de décrire les tâches affectées aux blocs fonctionnels. Il est possible qu'un nombre de logiciels soit nécessaire pour créer l'architecture NFV et la faire fonctionner. Une partie du travail de l'ETSI ISG NFV consiste à spécifier le transfert d'informations entre les différentes composantes de l'architecture pour répondre aux besoins de l'industrie et de développements d'outils appropriés. C'est peut-être le nombre de logiciels disponibles qui font émettre des doutes à [9] sur l'existence d'une plateforme MANO selon les normes ETSI.

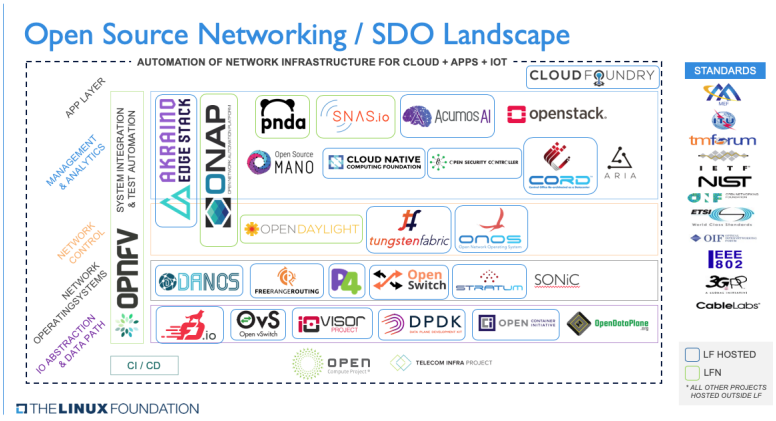

FIGURE 5.5 – Ensemble de logiciels open-source dans le cadre du NFV, mars 2020

5.3 Le placement des VNF

La section 5.1 a donné une définition de la SFC. Elle a décrit la SFC comme étant un ensemble ordonné de services de réseau abstraits sur lequel un ensemble de règles est appliqué. La section 5.2 a décrit une architecture sur laquelle la SFC peut être déployée et gérée. Le but du placement des VNF est de déployer des machines virtuelles et de les connecter virtuellement entre elles selon un ordre déterminé sur un ensemble de machines physiques interconnectées entre elles sur un réseau physique, de manière optimale, sans perturber les autres activités en cours et à moindre coût.

Sauf que le placement des VNF n'est pas un problème simple à résoudre. Il faut prendre en compte toute une série de contraintes afin de trouver la meilleure solution possible en fonction d'un objectif de placement déterminé [5]. Dans ce premier exemple de la figure 5.6, trouvé dans [4], la logique de la SFC est respectée mais le placement n'est pas optimal. La SFC est composée de 3 éléments ordonnés comme ceci : {Office → Firewall → Proxy → Business Logic → Sortie }. Pour les besoins de l'exemple, chaque lien entre les routeurs est considéré comme ayant les mêmes caractéristiques. Même si cela est possible, le placement en *a)* n'est pas optimal en comparaison avec *b)*. Le placement des fonctions pose également question. Celui du firewall devrait être entre le **Business Logic** et le *Regional Offices*. En *a)*, le trafic passe par le routeur où se situe la partie **Business Logic**. Cela présente un risque de sécurité puisque le flux traversant le routeur R1 n'est pas filtré par le firewall. Il est imaginable qu'une règle de sécurité impose le passage du flux réseau à travers un firewall avant de transiter sur le reste du réseau. On constate aussi que 2 routeurs sont traversés plusieurs fois par le flux de la SFC. Une règle de gestion du trafic peut imposer un nombre de sauts minimal ou une règle business peut demander une latence minimale. Sur le plan logique, le placement est correct (firewall puis proxy puis business logic) ; sur le plan physique, le placement optimal n'est pas rencontré. La proposition *b)* paraît plus optimale : l'ordre de la SFC est respecté, le firewall est traversé en premier, le trafic ne passe qu'une fois par chaque routeur ; le chemin est le plus court en terme de sauts. Les règles permettent de faire un choix dans les différentes possibilités de placements.

Pour définir et placer l'ensemble des VNF composant une SFC dans un environnement NFVI, le calcul du placement doit prendre en compte plusieurs éléments pour répondre adéquatement à la demande de service reçue. La complexité du calcul augmente au fur et à mesure que l'environnement NFVI s'élargit et se situe géographiquement à des endroits différents. Sur l'exemple précédent, si le nombre de routeurs augmente avec la possibilité d'y connecter d'autres services, le nombre de possibilités de placement va également croître. En ajoutant à l'équation des concepts comme le edge-computing, le cloud-computing ou la 5G/6G, la difficulté de trouver une solution optimale au placement s'accroît également. Dans sa conclusion, [5] affirme que l'objectif de placement souhaité oriente les décisions de placement. Parmi les objectifs cités, il y a le débit de données résiduel, la latence et le nombre de noeuds de réseau utilisés. Tiré du même article [5], le second exemple de la figure 5.7 montre que les règles peuvent également avoir un impact sur le déploiement

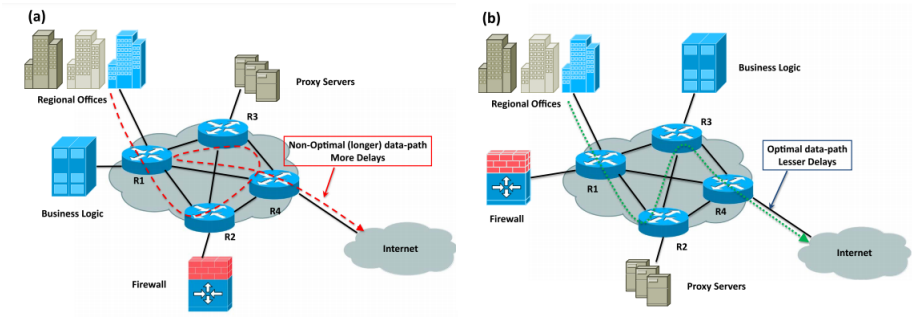

FIGURE 5.6 – Exemple d'optimalisation de placement, de [4]

de VNF supplémentaires pour les satisfaire.

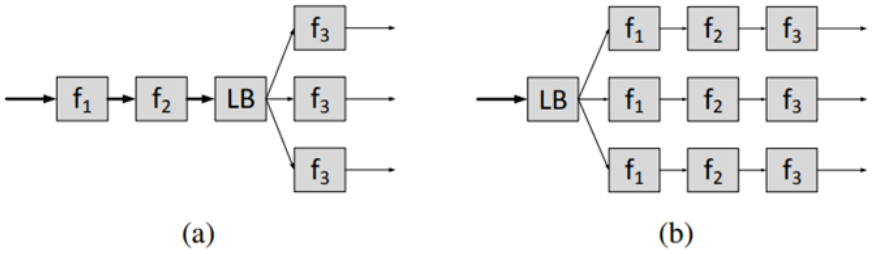

FIGURE 5.7 – Exemple de placement dans [5]

LB ici désigne un Load Balancer. Il répartit le flux entrant sur les 3 branches afin d'équilibrer le trafic sur les 3 branches. Pour être complet par rapport à [5], le débit sortant d'une VNF f_x est le même que le débit entrant. Comme le trafic entrant dans f_1 de (a) est plus important que celui entrant dans 1 des 3 instances de f_1 de (b), l'exigence de traitement sera plus importante pour f_1 de (a) que pour chaque instance de f_1 de (b) afin de maintenir la condition *débit entrant = débit sortant*. Nous avons donc une seule instance f_1 et f_2 de (a) dotées de plus de ressources pour effectuer la tâche demandée que du côté (b). Il est probable que la somme des ressources nécessaires pour les instances de (b) soient plus importantes qu'en (a). On notera également que le service offert par la partie (b) de la figure 5.7 est plus tolérant à la panne d'une des VNF f_1 ou f_2, ce qui peut être aussi un objectif. Le point faible commun aux 2 possibilités reste le LB.

Une autre méthode qui ajoute des VNF à la SFC est la décomposition d'un service demandé en un assemblage de services élémentaires ([23]). L'exemple donné dans l'article [11] est le cas du contrôle parental. Celui-ci peut être décomposé par une sous-chaîne composée d'une classification de trafic, suivi d'un proxy web puis d'un firewall. Mais cette méthode nécessite un calcul supplémentaire pour décomposer la requête initiale en plusieurs solutions et vérifier que le graphe des services décomposés est bien connecté et sans cycle.

La recherche de fiabilité est également génératrice de VNF supplémentaires ([23, 38]) pour pallier une panne d'une des VNF de la SFC ou du matériel de la NVFI. Un placement tenant en compte de cet aspect peut orienter le placement des VNF sur des serveurs différents ([21, 23, 38]), calculer un chemin réseau différent jugé plus fiable ([21]), anticiper une défaillance prévisible en déplaçant la VNF ([37]) . Le prise en compte de cet aspect est, par contre, en opposition avec la minimisation des VNF, l'optimisation des ressources et des coûts d'énergie.

Les axes de recherche décrits dans plusieurs documents de la bibliographie poursuivent plusieurs buts et objectifs comme la diminution des coûts techniques, opérationnels ou financiers, l'amélioration de la fiabilité, le respect des SLA/SLO et la diminution de la consommation énergétique. Associées à ces objectifs, des méthodes de placement différentes sont proposées. Pour appréhender la complexité du placement qu'induisent ces objectifs, un cas trivial est présenté sur le graphe 5.8 pour ensuite étendre la réflexion. Le graphe s'inspire de [39].

Postulons pour l'exemple que les 4 serveurs sont équivalents en terme de ressources allouées à la virtualisation et que les 3 VNF consomment chacune 30% de ces ressources.

Si le calcul place VNF1, VNF2 et VNF3 sur le serveur 1, on constate l'utilisation de 90% des ressources disponibles

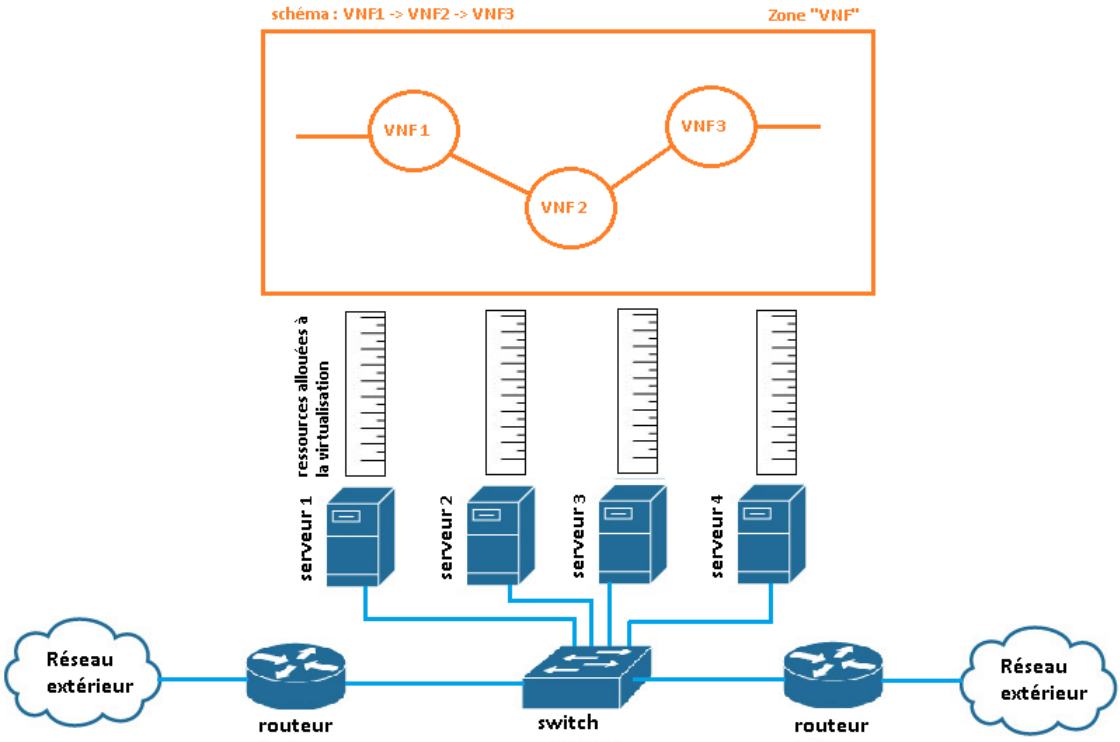

FIGURE 5.8 – Réflexions sur le placement

sur ce serveur. Les transferts de données entre VNF1 et VNF2 puis entre VNF2 et VNF3 se font d'une manière plus efficace car ces transferts ne passent pas par le switch physique. Dans [6], les auteurs signalent qu'un serveur inactif consomme jusqu'à 60% de sa puissance de pointe, chiffre datant de 2009. Dans ce cas de figure, les serveurs 2, 3 et 4 seraient éteints. Cette configuration consommant peu d'énergie par rapport au "tout sous tension", la rentabilité est optimisée du point de vue consommation électrique et, avec l'objectif d'améliorer le QoS grâce à une efficacité meilleure de la communication. Néanmoins, ce placement expose la SFC à plusieurs problèmes potentiels. Si le serveur 1 tombe en panne, il n'y a pas de serveur disponible immédiatement pour récupérer les 3 VNF en cours d'utilisation. Ainsi, le temps nécessaire au démarrage d'un autre serveur et le transfert des machines virtuelles vers celui-ci est une période d'indisponibilité de la SFC, potentiellement une violation du SLA et une perte des revenus. Toujours dans cette configuration, si la requête initiale est mise à jour et nécessite l'ajout d'une 4ème VNF, il faudra également un certain temps pour démarrer un nouveau serveur et le placement d'une nouvelle VNF. A partir du moment où l'on ajoute une nouvelle VNF, il est envisageable que le calcul du placement donne comme solution 2 VNF sur un premier serveur et les 2 autres sur un second serveur ainsi qu'une reconfiguration du trafic entre les VNF. Un autre cas similaire est la mise à jour de la requête initiale qui nécessite une augmentation d'utilisation de ressources et qui rend la configuration initiale obsolète (par exemple, la VNF2 nécessiterait 50% des ressources disponibles sur un serveur). Cette autre cause relance un nouveau calcul de placement et un nouveau déplacement de VNF.

Une autre configuration est le placement de VNF1 sur le serveur 1, VNF2 sur le serveur 2 et VNF 3 sur le serveur 3. La consommation d'énergie est globalement plus importante mais le matériel acheté est utilisé et participe à la génération de revenus. L'augmentation de capacité des VNF ne nécessite pas forcément un nouveau calcul de placement. Cela résoudrait en partie seulement des cas problématiques évoqués dans la configuration précédente. En effet, si un serveur tombe en panne, le 4ème serveur étant démarré, seul le transfert de la VNF est à gérer. Néanmoins, tout le trafic va systématiquement traverser le switch physique et cela peut poser des problèmes de performance du point de vue réseau. Les 2 configurations précédentes se sont faites à nombre de VNF constant au sein d'une même infrastructure. Une configuration intermédiaire avec 2 VNF sur un serveur et 1 VNF sur un autre pourrait également être envisagée mais le point important abordé ici est le choix nécessaire imposé par des règles et les implications du placement démontrées dans [5].

L'environnement NFVI a été conçu pour prendre en charge plusieurs SFC. Une fois une requête prise en compte et instanciée dans la NFVI, une autre va suivre. Chaque nouvelle requête complique le calcul du placement de VNF car il faut tenir compte des ressources déjà réservées, de l'état des liaisons avant de surcharger le tout avec une nouvelle demande. Dans [6], les auteurs nous proposent une solution qui consiste à utiliser les mêmes VNF quand c'est possible et à prioriser le trafic entre les mêmes VNF. [5] fait remarquer que certaines fonctions réseaux, comme un antivirus, ont une

configuration valable pour plusieurs SFC alors que d'autres fonctions doivent être paramétrées spécifiquement à chaque SFC. Cela donne schématiquement la figure 5.9 où le calcul du placement fera un choix entre le nouveau service prioritaire sur l'existant, le nouveau service secondaire sur l'existant ou bien la nouvelle instance est indépendante de l'instance existante.

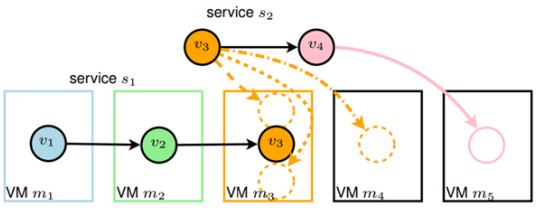

FIGURE 5.9 – Ajout d'une SFC avec priorisation, de [6]

Les différents exemples évoqués jusqu'ici montrent que les règles de gestion traduisent des objectifs business ou opérationnels et orientent le choix optimisé du placement des VNF. Les auteurs de la littérature reprise dans la méthodologie (section 6.6) ont envisagé des scénarios dans lesquelles ils étudient des solutions à ces objectifs. Une des pistes envisagées pour répondre à l'objectif de diminuer le coût opérationnel est de permettre à 2 SFC présentes dans le même environnement et ayant des VNF communes d'utiliser la même instance, quitte à augmenter les ressources nécessaires de cette instance pour prendre en charge les contraintes des 2 SFC [6, 25, 30, 34, 37, 38, 42]. Une solution pour améliorer la latence de traitement est le traitement en parallèle des paquets entre différentes VNF qui, pour être efficace, se fait sur le même serveur [20] mais aussi pour fiabiliser une SFC à la résistance aux pannes, [29, 38]. Pour améliorer la fiabilité des réseaux, [21] propose une solution utilisant des noeuds de secours, comme [23] dans un scénario mobile. La performance du réseau est aussi une piste d'amélioration étudiée [28, 43]. Pour tenir compte du placement avec le système en cours de fonctionnement, les travaux comme [8, 25, 28, 39, 44, 45] ont intégré ce paramètre dans leur développement. Alors qu'il est quasi impossible de déterminer le moment où une nouvelle requête va arriver, la prédiction de l'arrivée de ces dernières est étudiée dans [40, 41, 44, 46]. Enfin, nombreux sont les travaux à prendre la dimension coût/revenu dans leur réflexion, voire même basent leur réflexion de manière centrale sur ce sujet [7]. Le graphe 5.10 issu de [7] nous présente la balance entre les utilisateurs et les fournisseurs de services "cloud".

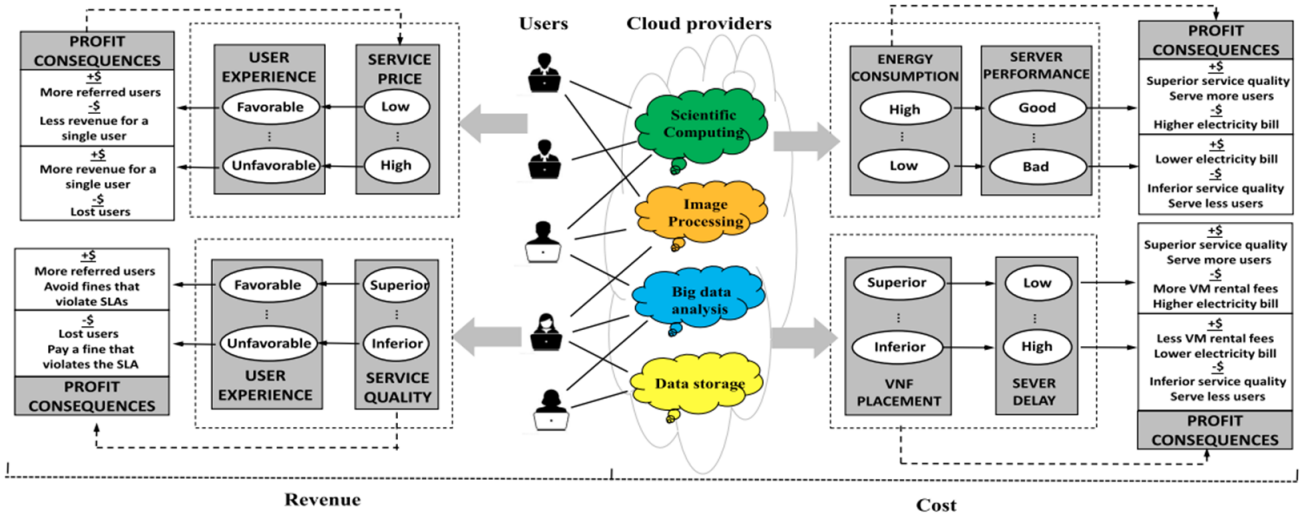

FIGURE 5.10 – Balance revenu/coût pour les fournisseurs de services "cloud" [7]

5.4 La complication de l'environnement

Les règles, dictées par des objectifs de différentes natures ont un impact sur le placement des VNF. D'autres choix ont été faits pour rencontrer ces mêmes objectifs. Ces choix modifient l'environnement technologique dans lequel s'insère la NFV. Cela agrandit les possibilités de placement et complique grandement le calcul du placement des VNF. De plus, ces nouveaux arrivants que sont la 5G, le découpage de réseaux (*Slicing*), l'informatique en nuage (*Cloud*) et l'informatique

en périphérie (*Edge*) viennent avec leurs architectures, leurs buts et leurs objectifs.

La 5G a, par concept, des exigences de performances [19] sur lesquelles de nouvelles applications s'appuieront pour pouvoir être développées et certains travaux prennent en compte cette technologie [8, 36, 47, 48]. Elle a sa propre architecture [6], comme celle définie par 5G-PPP (graphe 5.11).

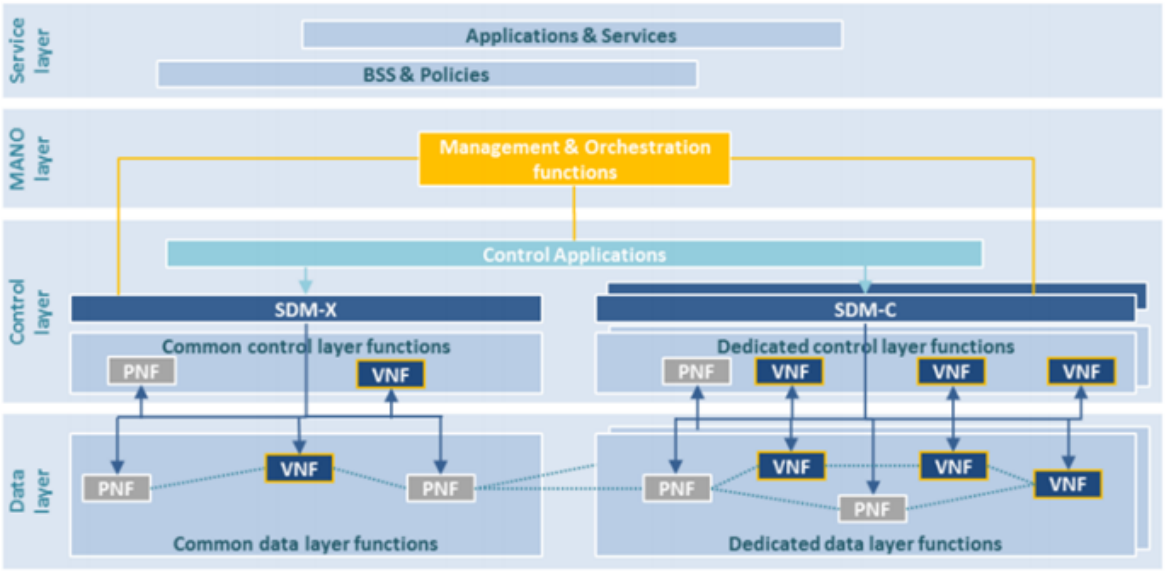

FIGURE 5.11 – Vue architecturale de la 5G selon 5G-PPP, de [6]

La 5G vient aussi avec le découpage de réseaux (*network slicing*). Cette architecture de réseau autorise le multiplexage de réseaux logiques et virtualisés. Ces réseaux, bien que sur la même infrastructure physique, sont isolés les uns des autres, répondent à des niveaux de service différents pour des applications particulières. Cette classification du trafic réseau en fonction de son application rappelle la fonction de classification expliquée dans [1]. Le graphe 5.12 décrit le concept de découpage réseaux pour la 5G.

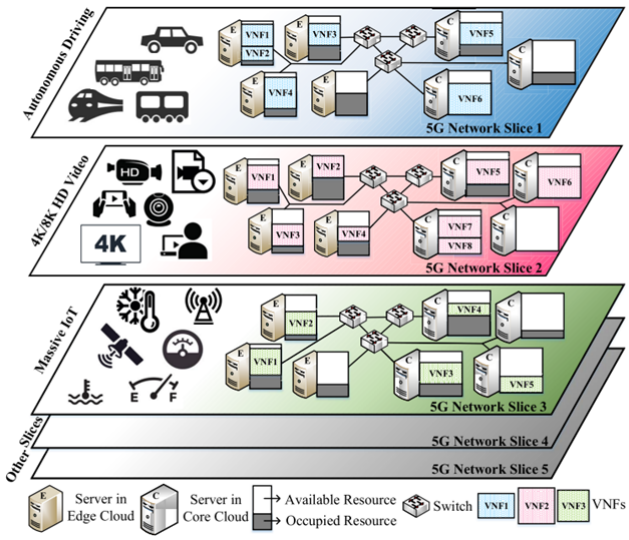

FIGURE 5.12 – Tranches réseaux dans la 5G, de [8]

La prise en compte de l'informatique en nuage, *"cloud-computing"*, est étudiée dans [7, 36, 45]. Le "cloud" permet la connexion à un service hébergé sur des serveurs "quelque part", ce quelque part étant l'infrastructure d'un fournisseur de service. Ici, les objectifs rencontrés sont la réduction des coûts de l'IT et le report du risque sur le fournisseur en cas de problème avec le service utilisé via un SLA. Mais cela induit la gestion du trafic entre le client et le fournisseur du service

avec les VNF nécessaires et des contraintes additionnelles pour le bon fonctionnement de cette liaison.

Le cas de l'informatique en périphérie, *"edge-computing"*, est étudiée dans [8, 9, 35, 37, 40, 41]. Cette méthode d'optimisation rencontre le besoin d'une latence plus faible pour certains services. Les serveurs edge-computing étant placés au plus près des utilisateurs, réduisent la distance parcourue par le trafic vers ces serveurs et diminuent la bande passante utilisée globalement sur le réseau.

La figure 5.13 permet d'avoir une vue macroscopique de l'enchevêtrement des *cloud-computing, edge-computing, 5G et slicing*. On comprend rapidement que le nombre d'endroits où une VNF peut être placée a considérablement augmenté, donnant un nombre important de possibilités et, par conséquence, a un impact sur le calcul du placement, par exemple dans [27].

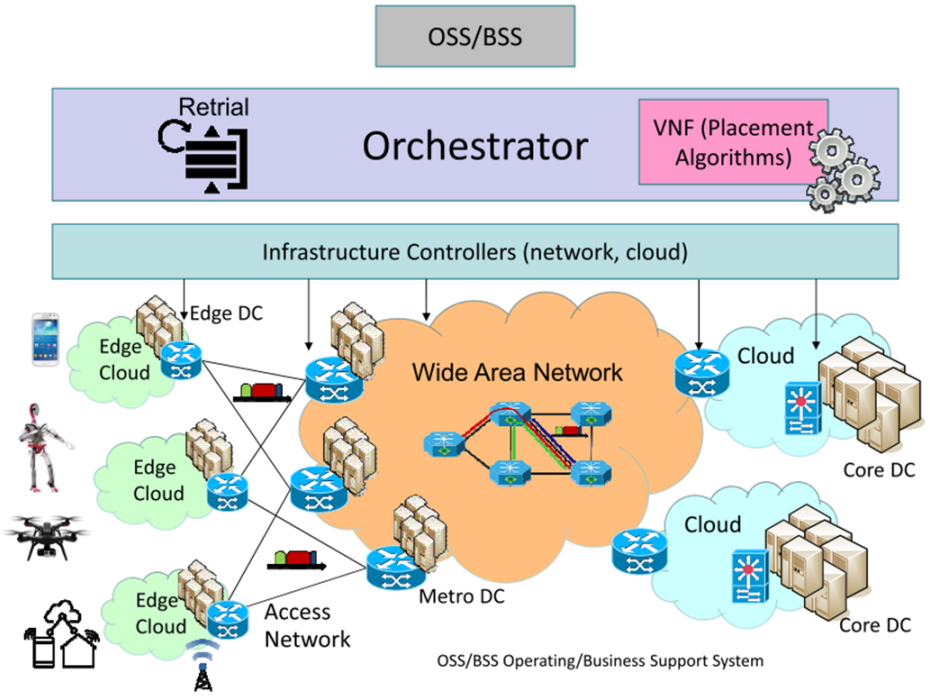

FIGURE 5.13 – Vue macroscopique de la NFV et des différents réseaux, de [9]

Pour compléter ces cas généralistes, il y a des cas spécifiques qui viennent se greffer. Des domaines d'applications spécifiques comme l'étude océanographique ou les services embarqués sur drone [13, 37, 49]. Ces environnements amènent leur lot de spécificités et de contraintes comme la dérive des bouées ou l'énergie restante des batteries pour les drones, ce qui va entraîner un nouveau calcul du placement des VNF dans le réseau.

D'autres études utilisent de nouvelles méthodes comme le "Machine Learning" pour répondre à différents scénarios [7, 13, 40, 46, 50]. Les auteurs parlent d'ajouter des agents sur le réseau, nécessaires au bon fonctionnement de leurs méthodes mais qui génèrent également du trafic et ont un impact sur les ressources réseaux allouées au trafic réseau des SFC.

5.5 L'impact du déploiement des instances VNF

La section 5.4 montre un des aspects de la difficulté du calcul du placement optimal des VNF sur l'infrastructure physique. Un autre aspect qui impacte le placement est le déploiement des VNF. Selon [7], *le lancement d'une nouvelle instance d'une VNF entraîne le transfert d'une image de machine virtuelle vers un nouveau serveur.* Dans le cas de l'article, l'image étant placée sur une infrastructure cloud, il faut un certain temps pour la récupérer, la mettre sur le serveur visé et l'intégrer à l'ensemble. Ce téléchargement utilise le réseau pour transiter, utilisant de la bande passante utile mais nécessaire pour effectuer l'opération. Cette utilisation de ressources du réseau physique allouée à la SFC et le coût fixe de l'instanciation des VNF est appelé **coût de mappage** [27]. Cependant, un nombre trop grand de transferts peut avoir un impact sur l'ensemble du trafic réseau en diminuant la bande passante disponible pour les SFC. Cela a aussi pour effet

d'augmenter les coûts de transmission puisqu'il y a une augmentation du trafic et, éventuellement, une compensation de cette augmentation de trafic par une augmentation de capacité des lignes. Les règles de placement des VNF doivent également prendre en compte ce problème de déplacement de VNF. La recherche de la minimisation du nombre de noeuds utilisés est une des pistes envisagées pour répondre à ce problème [5, 10].

5.6 Les liaisons entre les VNF

Jusqu'à présent, seul le placement des VNF a été passé en revue. La mise en contexte des évolutions de l'environnement de la NFV ont permis de se faire une idée de la complexité du placement. Mais une autre composante de la problématique de placement réside dans les liaisons reliant les VNF. Si l'on considère la figure 5.14, les liens entre les fonctions réseaux sont des liaisons qui, in fine, reposent sur un réseau physique. On peut s'en convaincre sur la figure 5.8. Le réseau a son lot de contraintes physiques et de problèmes à résoudre [5]. Il faut donc prendre en compte l'état du réseau [24, 43] et calculer le meilleur chemin possible tout en respectant les attendus commerciaux. Cela revient à un problème de routage des données et le fait que plusieurs auteurs [6,8,9,11,13,20,21,23,26,27,34,35,45,49,50] joignent le SDN au VNF, ce qui n'est pas anodin. Le SDN est un paradigme où le plan de données est décorrélé du plan de contrôle. Ce plan de contrôle gère le calcul des routes en centralisant l'état des matériels type routeurs, switches. Ainsi, avec un centre de décision central, les mises à jour des réseaux sont généralisées.

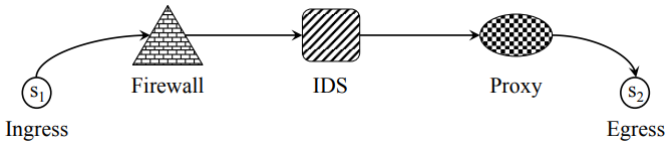

FIGURE 5.14 – Schéma d'une SFC, de [10]

Un aspect évident à prendre en compte est la congestion du trafic. Pour faire une Lapalissade, il est inutile d'essayer de faire transiter un trafic de 1 Gbps sur une liaison physique qui est prévue pour un trafic de 100 Mbps. Si trop de trafic de SFC différentes passent par la même ligne, il y a concurrence entre les différents trafics et cela impactera le QoS/SLA attendu pour chaque SFC. Sur l'autre plateau de la balance se trouve le coût des lignes. Trop de lignes peu utilisées mais garantissant les SFC vont augmenter les frais des opérations. Dans un environnement concurrentiel [7], notamment pour les services "cloud", cela est loin d'être négligeable.

La latence est également un des facteurs à étudier. Dans la recherche du meilleur chemin possible, avoir le temps nécessaire au transfert d'un paquet sur le réseau le plus bas est un critère prioritaire pour certaines catégories d'application et moindre pour d'autres. Avec la superposition des trafics, il faut envisager d'utiliser d'autres chemins, peut-être plus longs mais qui remplissent les SLA/QoS des applications. Le volume des données transférées est en augmentation, notamment avec l'Internet des Objets (*IoT*) [40, 50] ; la latence tend également à la hausse ([50]).

5.7 La représentation du réseau physique

Pour permettre le calcul du placement optimal des VNF, il convient de traduire en langage mathématique le réseau physique. La représentation faite par les auteurs de [5] transpose le support physique comme un graphe orienté $G = (V, E)$ où V est l'ensemble des noeuds physiques et E est l'ensemble des liaisons. Les noeuds ont des attributs reprenant la capacité de calcul, de mémoire et de stockage ; les liaisons ont des attributs de latence et de débit. Sur ce graphe G, connecté et orienté, on fait correspondre la chaîne de VNF avec les VNF sur les noeuds et les liens entre les VNF sur les liaisons physiques, comme le montre la figure 5.15. Les articles [9, 20, 22, 24, 26, 27, 34, 39, 48] ont la même approche pour le graphe du réseau physique. [24] précise que son graphe représente un large réseau géographiquement distribué, typiquement celui d'un opérateur de télécommunications. [9] envisage également un cadre d'opérateurs de télécommunication où la pondération des liens prend en compte la distance entre deux noeuds et la bande passante disponible. [26] traitant du placement des VNF dans un réseau SDN, le graphe dirigé et acyclique porte sur le plan de données (Data Plane). Le graphe utilisé par [20] décrit le réseau d'un centre de données.

Si la représentation du graphe de l'infrastructure physique varie peu, certaines différences sont notables dans d'autres articles. Si les noeuds et les liaisons sont sensiblement proches du point de vue des attributs, [8, 10–13, 43] considèrent

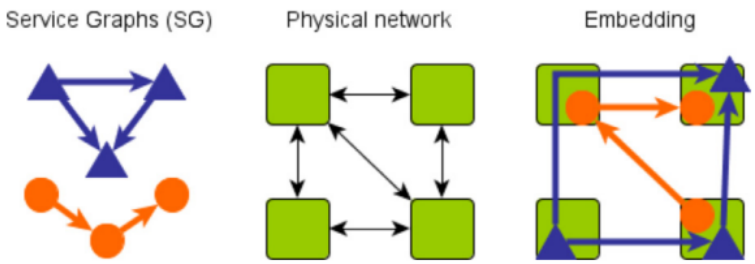

FIGURE 5.15 – Placement des VNF sur une infrastructure physique, de [11]

que le graphe de l'environnement physique est non dirigé.

Dans leurs articles respectifs, [21, 23] considèrent le réseau physique comme étant un ensemble de machines physiques interconnectées par un ensemble de liens. On ne retrouve pas explicitement la notion de graphe de réseau physique sur lequel on applique le graphe des VNF. [44] conçoit les ressources physiques comme deux ensembles. Le premier ensemble se compose de serveurs hétérogènes et le second reprend pour chaque liaison entre deux serveurs des attributs de coûts de transferts et de nombre de sauts. [46] traite de *deep-learning* et voit la topologie du réseau physique comme un réseau convolutif.

5.8 La formulation du problème du placement des VNF

La partie physique ayant un modèle de représentation, il faut passer à la formulation du placement optimal des VNF. Les sections précédentes montrent que c'est une opération complexe. Les contraintes applicables sont nombreuses et suivent des logiques différentes. Un objectif remonte quasi systématiquement dans la littérature de la section 6.6. Les fournisseurs de services recherchent la rentabilité de leur système ; leurs clients demandent des services fiables au meilleur prix. Pour rencontrer ces attentes, de nombreuses pistes ont été explorées et reprises dans la documentation de la méthodologie. Des travaux, soutenus par des institutions publiques ou des entreprises, ont amené leurs solutions et montré leurs limites. Certaines limites ont été surpassées mais il reste certaines questions en suspens.

La complexité du problème a été évoquée dans vingt-deux articles ([5–11,20,22,24–26,28–30,34,36,39,40,43,47,48]) comme étant NP-Difficile (*NP-Hard*). Cette complexité a été soit démontrée dans les articles, soit citée par référence. Elle est la conséquence de la manière dont les auteurs des articles ont posé le problème du placement des VNF. De nombreux auteurs ([7, 10–12, 20–23, 28, 37, 39, 41, 45, 48, 51]) ont présenté une solution d'optimisation comme étant un programme linéaire en nombres entiers (*ILP*) ou un programme linéaire en nombres entiers mixtes (*MILP*). Cette approche est, selon le texte *0-1 Integer Programming* de Richard M. Karp, un problème NP-Difficile.

Des tentatives de résolutions du problème du placement des VNF ont été proposés et des fonctions objectives établies dans certains travaux ont pu être codées et testées dans des simulateurs. Par exemple, dans [12] en 2014, les auteurs montrent point par point la construction de leur fonction objective. Une fois la fonction codée pour créer un algorithme, le programme est employé par CPLEX pour fournir un résultat. Celui-ci sera fourni après un temps d'exécution de l'algorithme de 16 secondes maximum dans le cadre de leurs tests. Les calculs ont été effectués sur une machine Linux équipée d'un Core I3 et 4 GB de RAM. Le réseau servant de support de tests est considéré comme petit par les auteurs. Selon leur diagramme de performance (figure 5.16), le temps d'exécution semble linéaire en fonction du nombre de requêtes de chaînes de services.

Bien qu'attendu par le fait que cela soit un problème NP-Difficile, la pierre d'achoppement des solutions basées sur un algorithme ILP est clairement identifiée par les chercheurs : le temps de réponse de ce dernier pour fournir une solution. Ce problème a été soulevé notamment dans [11] où les auteurs ont fait l'exercice de développer un algorithme ILP puis un algorithme heuristique et les ont comparés entre eux. Pour tester ces algorithmes, les auteurs ont pris la topologie de "British Telecom Europe" composée de 24 noeuds et de 37 liens pour le petit réseau et la topologie "Interoute" comptant 110 noeuds et 148 liens pour le grand réseau. Le diagramme figure 5.17, montre le temps nécessaire à l'algorithme pour placer des chaînes de services de fonctions contenants 5 et 10 services réseaux dans chacune des topologies. Les auteurs informent que le temps d'exécution pour l'algorithme heuristique DSBM-10 (Decomposition selection-backtracking mapping avec 10 NF) ne dépasse pas les quelques 100 ms.

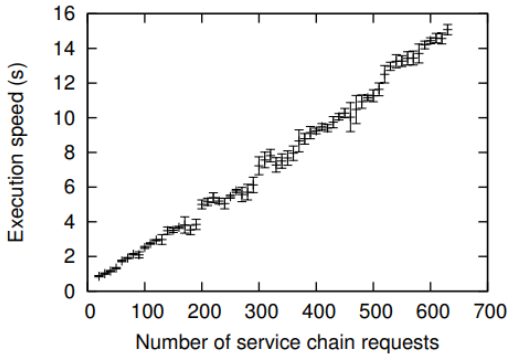

FIGURE 5.16 – Diagramme de performance d'un algorithme ILP, de [12]

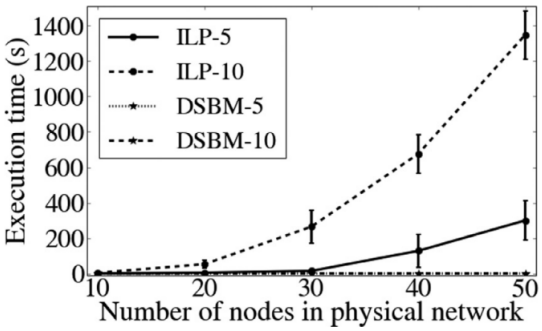

FIGURE 5.17 – Temps d'exécution de l'ILP et de DSBM pour des graphes de services avec 5 et 10 fonctions de services. L'intervalle de confiance à 95 % des valeurs moyennes rapportées est indiqué. Source : [11]

Il faut se prémunir de comparer les figures 5.16 et 5.17 tant les scénarios autour desquels sont construits les algorithmes et les tests sont différents. On constate sur le graphique extrait de [11] que la taille de la topologie et du nombre de fonctions de réseaux à placer sont des facteurs de performance. Le prix à payer par l'algorithme heuristique est pour les auteurs de [11] une solution moins efficace par rapport au placement optimum déterminé par l'algorithme ILP.

Avec un autre scénario envisagé et une méthodologie semblable à celle de [11], [10] retient également ce choix de solution moins efficace mais plus rapidement obtenue. Le tableau de comparaison entre leur algorithme ILP traité par CPLEX et leur heuristique donne des résultats similaires. Bien que leur solution ne soit pas optimale, les auteurs concluent notamment à une baisse du coût d'exploitation du réseau d'un facteur 4.

Topology	CPLEX	Heuristic
Internet2 (12 nodes, 15 links)	34.99s	0.535s
Data Center (23 nodes, 43 links)	1595.12s	0.442s
AS-3967 (79 nodes, 147 links)	∞	2.54s

TABLE 5.1 – Temps d'exécution moyen entre CPLEX et une heuristique, de [10]

L'étude de la documentation récupérée (chapitre 6.6) montre que d'autres articles ont également procédé de la même manière : formulation ILP du placement, algorithme ILP, algorithme heuristique et comparaison des résultats. Du point de vue du temps d'exécution, l'algorithme heuristique est plus performant. Ainsi démontré, les algorithmes ILP ne sont pas une solution pour de grands réseaux [10–12, 26]. Les chercheurs ont intégré la méthodologie suivante : mathématiquement, le problème d'optimisation du placement est un ILP/MILP ; sa solution donne un algorithme heuristique. Les problèmes MILP autorisent pour les variables d'autres types de valeurs que les valeurs entières. Ainsi, les différentes recherchent obtiennent une solution proche de celle fournie par un algorithme (M)ILP, dans un temps raisonnable, fonction de la taille du réseau. Cette constatation a été confirmée récemment par les auteurs de [52].

5.9 Les solutions heuristiques

Les solutions optimales trouvées par un algorithme (M)ILP ne passant pas l'écueil du passage à l'échelle, des solutions sur base d'algorithmes heuristiques obtiennent des résultats sous-optimaux proches de la solution optimale dans les articles qui font la comparaison entre ces deux types d'algorithme, comme dans [10]. Le graphe 6.6 catégorise les différents articles de la méthodologie 6.6 en différents groupes. Ainsi, les articles en jaune, orange et vert sont des articles présentant des solutions heuristiques au problème de placement. Ils représentent la majorité des articles trouvés.

Selon [7], la capacité à résoudre le problème NP-Hard de l'optimisation du placement des VNF, l'optimisation des coûts et des performances devient avec les solutions heuristiques le principal sujet de recherche. Par ailleurs, la démonstration a été faite d'une meilleure efficacité des solutions heuristiques pour obtenir rapidement une solution proche de l'optimale (voir section 5.8). Des études se sont attachées à la résolution du problème d'optimisation du placement en mettant en avant certaines propriétés, environnement, scénario et/ou objectifs spécifiques. Les titres de certains articles parlent d'eux-même : *Reliability-Aware Service Provisioning in NFV-enabled Enterprise Datacenter Networks [21]* qui traite du placement des VNF dans des réseaux de centres de données NFV en fonction de la fiabilité ; *Online Scaling of NFV Service Chains across Geo-distributed Datacenters [25]* dont l'objet est le placement en ligne des VNF dans des centres de données géodistribués. Ces deux exemples illustrent le propos mais il y a d'autres titres dans les références qui se prêtent à cet exercice comme [27, 28, 41, 47] pour ne citer qu'eux.

L'objectif quasi systématique de ce type d'études tourne autour d'une amélioration des coûts ou de revenus, que ce soit via la réduction de la consommation d'énergie, la minimisation du coût de déploiement des implémentations des services, l'utilisation optimale de la bande passante, la prise en compte d'un maximum de requêtes, la robustesse de la SFC. Ces objectifs s'appliquent à des scénarios rendus plus complexes par l'environnement technologique comme le cloud-computing et le edge-computing, la 5G. Deux autres aspects ont aussi été explorés par les solutions heuristiques : le caractère online/offline ([8,25,28,44,45]) de l'application de l'algorithme et le caractère prédictif ([7,22,44]de celui-ci.

Le caractère "online" dans la documentation trouvée représente l'application de l'algorithme sur de l'existant en cours de fonctionnement. Les versions "offline" calculent le placement de VNF. Une fois le résultat obtenu, les VNF sont placées puis mises en production. Le calcul "offline" se fait à un moment *t* avec un ensemble de variables, de caractéristiques fixées à ce moment *t*. L'environnement dans laquelle la nouvelle SFC est placée va varier au cours du temps. Il y aura plus ou moins de requêtes en entrée, il y aura suppression ou ajout de nouvelles SFC, il y aura des "pannes" de VNF ou de matériel, il y aura des changements de routes pour le trafic. Ces modifications dans le temps peuvent amorcer un nouveau calcul de placement des VNF existantes. La version "offline" n'est pas, par définition, très réactive à ces changements. Le temps qu'une modification soit mise "en ligne", une autre modification peut engendrer un nouveau calcul. Le but des algorithmes "online" est d'agir au plus près des modifications rencontrées et d'apporter des correctifs éventuels afin de respecter les règles et SLA/QoS.

Une requête peut arriver à n'importe quel moment et le nombre de requêtes reçues n'est pas constant dans le temps, il y a donc des pics et des creux qu'il serait intéressant d'anticiper. Le caractère prédictif est la prédiction du comportement des requêtes dans le temps. Une fois le comportement des requêtes connus, la possibilité est offerte de placer au préalable les composants des SFC au bon endroit et d'absorber un plus grand nombre de requêtes tout en garantissant la qualité de service. De cette manière, l'algorithme permet une meilleure gestion des ressources disponibles, soit à affecter, soit à redistribuer. Cela rend possible aussi le déploiement des VNF dans des périodes creuses pour le trafic. Cependant, une prédiction admet un certain nombre d'erreurs ([7, 22] et ces erreurs ne doivent pas mettre en péril les SFC déjà déployées et en cours d'exploitation.

Ces deux orientations dans la conception des algorithmes sont nécessaires pour prendre en compte les fluctuations des requêtes dans le temps. L'accroissement des requêtes provient du plus grand nombre d'applications utilisant les réseaux, comme l'IOT, mais en prévision de la 5G/6G et des nouvelles applications nécessitants une plus grande fiabilité, un plus grand volume de données transférées. Il faut aussi tenir compte des pics de requêtes à des moments donnés qu'il faut pouvoir absorber rapidement tout en n'y perdant pas en qualité de service. Une autre technologie, abordée dans [50] peut prendre en compte l'ensemble des critères et apporter des solutions de placements à la fois "online" et de manière prédictive.

5.10 Les solutions Machine-Learning

Les solutions type Machine-Learning (ML) sont également considérées comme étant une solution au placement des VNF. L'idée d'un plan de la connaissance qui remonterait à une entité centralisatrice des informations sur l'état du réseau

et de ces composants et qui permettrait, grâce à ces informations d'avoir une connaissance de l'architecture de l'ensemble est présentée par les auteurs de [50]. En reprenant les travaux de D. Clark (*"A Knowledge Plane for the Internet"*), ils proposent de centraliser la télémétrie du réseau en temps réel sur une plateforme d'analyse et, avec des techniques d'apprentissage en profondeur, d'exploiter les connaissances extraites de cette télémétrie pour contrôler et maintenir le réseau. Étendre ce principe permet de prendre en compte la complexité de l'environnement, l'application "online" des modifications et le caractère imprévisible des requêtes via entre autres l'apprentissage du comportement du réseau.

5.10.1 Le principe du machine-learning

Une explication simplifiée de l'apport du machine-learning au placement des VNF est utile. L'explication ci-dessous se base sur les articles présent dans la méthodologie (section 6.6) et se veut plus pédagogique qu'extrêmement précise concernant le sujet "machine-learning".

L'orchestrateur et ses composants sont repris par le *control panel* alors que le *data panel* reprend les différents composants virtuels de la SFC ainsi que la partie physique qui soutient ces composants virtualisés. Le plan *Intelligent algorithm plane* reprend les éléments du machine-learning. Le cadre de gauche représente la partie "statistiques et prédictions", le cadre de droite représente la partie "prise de décisions".

Des agents collectent et envoient des informations sur l'état des composants de la SFC à l'orchestration. Dans le cadre du machine-learning, ces informations sont envoyées en temps réel. Ces informations sont également envoyées à la partie "statistiques et prédictions". Sur cette base statistique, un modèle de l'ensemble du réseau est proposé sur lequel est joint une partie prédictive. Ce modèle est transmis au processus de prise de décisions. Celui-ci choisit le meilleur placement des VNF et transmet son résultat à l'orchestrateur qui, par rapport à l'existant, transmet les modifications à l'infrastructure, à savoir création de nouvelles VNF, déplacement des VNF existantes, ajustement des ressources des VNF existantes ou mise en sommeil voire suppression de VNF. Les chemins de routage peuvent également être modifiés. [13] démontre l'utilisation des techniques de ML dans le cas de l'orchestration des VNF. Le schéma 5.18 est tiré de leur document.

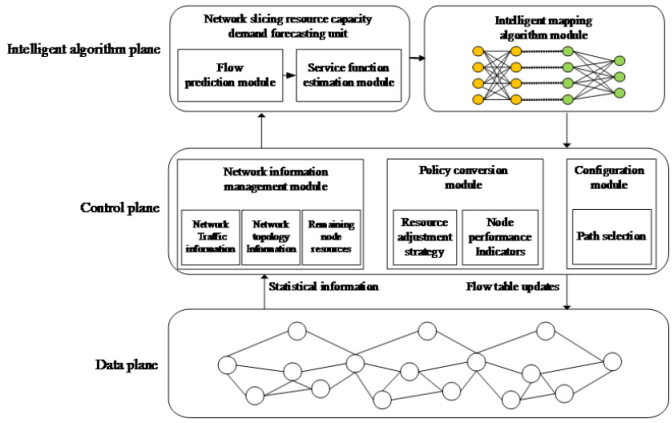

FIGURE 5.18 – Schéma présentant le fonctionnement de VCAD, de [13]

Quelques précisions sur le principe sont à apporter.

Les méthodes employées pour créer un modèle et prendre une décision sont appelés algorithmes stochastiques. Par définition, ces méthodes se basent sur les statistiques pour produire un résultat. Les méthodes sur lesquelles s'appuient ces algorithmes sont nombreuses (voir table 1 de [40], tables 4 et 5 de [35], table 1 de [46]).

Les données transmises en temps réel ont un poids sur le réseau et plus il y a des agents présents, plus la quantité de données gagne en importance. Il faut veiller à ce que ce poids n'impacte pas le trafic des SFC gérées par ce système. Ce sont également ces données qui, par accumulation dans le temps, vont permettre au modèle de converger vers une version optimale. Il faut donc une certaine quantité d'informations pour parvenir à un résultat correct. De plus, la convergence peut prendre du temps ([46]). Une technique pour apprendre à cette partie "statistiques et prédictions" le comportement de l'ensemble est de lui transmettre un jeu de données représentatives du comportement de l'ensemble. Ainsi, le temps de convergence diminuerait. Des données antérieures extraites de cet ensemble et couvrant une période suffisamment grande

pour être réaliste feraient un ensemble de référence pour cet ensemble. Comme c'est un modèle statistique, il est probable que des erreurs de prédictions l'affectent. Il faut veiller à ce que ces erreurs aient un impact minimal sur le bon fonctionnement des SFC en cours de productions et une balance des coûts/bénéfices doit se faire entre les changements apportés, leurs impacts sur le fonctionnement général et le respect de chaque SLA/QoS ([40]). Un autre point d'attention est la fréquence des changements à apporter. Il faut un certain temps pour que ces effets soient répercutés sur l'opérationnel et que les données remontent pour vérifier l'efficacité des changements apportés ([13]). Ici aussi une balance est nécessaire entre le temps pour vérifier l'efficacité du placement des VNF et le maintien d'un placement de VNF moins efficace.

5.10.2 Les articles traitants de machine-learning

Une étude [40] porte sur la fiabilité de l'approvisionnement en ressources pour les environnements "edge" et "cloud". Reconnaissant la complexité du problème, les auteurs mettent en avant l'apprentissage pour répondre à cette problématique et, pour atteindre leur but, décomposent le problème en trois catégories auxquelles sont attachés des techniques : la caractérisation et la prédiction de la charge, le placement et la consolidation des composants, l'élasticité et la remédiation des applications. Au final de leur étude, les auteurs observent qu'*"'en moyenne, les techniques d'apprentissage automatique donnent de meilleurs résultats que les méthodes traditionnelles, en particulier lorsqu'il s'agit d'environnements importants et complexes."*. Cependant, les auteurs soulèvent aussi une série de questions sur l'emploi et les conditions à remplir pour utiliser ces techniques de machine-learning. Dans leur étude [7] sur l'optimisation des profits dans un contexte de fournisseurs de solution cloud, les auteurs avancent également que les différentes techniques de machine-learning peuvent être une solution qui augmente les profits mais ne s'avancent pas plus sur le sujet.

L'approche de [39] vise à répondre au problème d'optimisation des performances sur le long terme. Constatant le comportement variable de l'arrivée des demandes et de grandes variations imprévisibles du trafic, les auteurs découpent le problème en plusieurs sous-problèmes et utilisent notamment des techniques de DRL ("Deep Reinforcement Learning") en tenant compte du QoS/QoE (Quality Of Experience) comme paramètres pour faciliter le placement des VNF.

La publication [13] cherche à améliorer la flexibilité de la fourniture de services en adaptant en ligne les ressources allouées aux VNF. En combinant plusieurs méthodes de machine-learning, les auteurs donnent un modèle de prédiction décrit comme précis à 98,31%, une méthode d'ajustement provenant du DRL. Les résultats donnés sont d'une amélioration du taux d'utilisation de 7,72% et d'une réduction de consommation d'énergie de 10,17% grâce à la mise hors service de certaines machines.

Alors que l'étude [35] affirme que les méthodes basées machine-learning vont être plus que nécessaires pour les réseaux intelligents du futur et leurs gestions quasi-complètement automatisées, le document trouvé le plus récent, [46] se place sur les réseaux 5G et futurs 6G. Les spécifications de haute disponibilité et de performances de ces nouveaux réseaux informatiques nécessitent des réactions rapides pour maintenir les services utilisant ceux-ci. La complexité de la gestion et de l'orchestration est aggravée par le découpage du réseau (slicing) et les solutions attendant une réaction humaine atteignent leurs limites. Aussi, l'utilisation du DRL et des réseaux neuronaux est envisagée pour le placement optimal des tranches et leur mise à l'échelle dans des conditions strictes de débit, de latence et de fiabilité imposées par l'utilisation de la 5G. La problématique rencontrée pour le placement et la mise à l'échelle des tranches de réseau est l'instanciation/reconfiguration de la tranche via l'allocation de ressources et le placement des VNF.

Chapitre 6

Critiques de l'état de l'art

6.1 Résumé de l'état de l'art

L'état de l'art a parcouru plusieurs dizaines de documents. Cette documentation permet de retracer l'évolution des algorithmes de placement des VNF tout en mettant en parallèle les évolutions technologiques des réseaux qui soutiennent le-dit placement des VNF. Les possibilités techniques offertes par la virtualisation ont ouvert la voie vers la softwarisation des différentes middleboxes physiques qui se situent sur le parcours des flux réseaux. Les concepteurs, notamment au niveau de l'ETSI, ont créé un modèle pour permettre l'enchaînement des middleboxes virtualisées. Pour gérer le cycle de vie de ces enchaînements, la partie orchestration du modèle a besoin de connaître :

— les ressources physiques dont elle dispose ainsi que la connaissance de leur état,
— d'un ensemble de règles pour faire fonctionner l'ensemble physique et logiciel,
— la connaissance des machines virtuelles créées et leur état,

A partir de ces connaissances, l'orchestrateur peut placer les VNF adéquatement sur l'existant. Le problème de l'optimisation du placement de ces machines virtuelles peut être résumé comme la tentative de faire correspondre de façon efficace et optimale une SFC virtuelle sur un réseau physique. Une partie de l'état de l'art a montré différents paramètres influençant le problème de placement des VNF. Les premières recherches évoquées dans ces articles ont démontré la faisabilité de cette gestion via un algorithme de placement de type (M)ILP. Elles ont également montré, comme le prévoyait l'étude mathématique, un problème d'efficacité de ces algorithmes ILP à obtenir rapidement une solution lorsque l'échelle commence à grandir. Ce problème de passage à l'échelle a été résolu par le développement d'heuristiques algorithmiques qui obtiennent des solutions sous-optimales des solutions parfaites des ILP.

Le contexte technologique a évolué sur la période de temps que couvre la recherche documentaire (section 6.6). Le chapitre 5 pointe l'explosion des requêtes à traiter (nouvelles applications, IOT), l'augmentation des volumes de données transférées, la complication des réseaux (Edge-computing, cloud-computing, 5G, 6G). "Plus d'appareils plus consommateurs de ressources réseaux sur des réseaux plus fiables et performants". A ce contexte s'ajoute l'évolution dans la gestion de la SFC comme le résultat "en ligne" de l'algorithme de placement, la prédiction du comportement du réseau et son corollaire l'adaptation "en ligne" des ressources des VNF. Au vu de ces évolutions, des études mettent en avant une solution de type Machine-Learning pour prévoir le comportement du trafic réseau et anticiper la mise en place des VNF.

Sur les 38 documents trouvés (Section 6.6), 34 présentent une solution à l'optimisation du placement et chiffrent les gains que leur solution apporte. Seuls 4 articles ([7, 35, 46, 50]) font une synthèse de solutions exposées selon une approche particulière. Les gains annoncés ont été calculés sur base de simulations. Certaines simulations prenant en compte la topologie physique, sont annoncées proches de la réalité. D'autres ont pris le parti de créer leur environnement de tests sur base d'un scénario. A la lecture des articles, il est possible de construire le tableau synthétique 6.1 qui reprend le nombre de noeuds/arêtes des réseaux et/ou le modèle employé. Les informations composant ce tableau sont reprises lorsque le document le spécifie clairement ou sont rapidement calculables ; le " ?" étant à interpréter dans le sens "non fourni".

6.2 Étude de la figure 6.6

L'état de l'art a été réalisé via des articles trouvés sur base d'une requête de mots-clés et d'articles reçus de mon promoteur traitant de ce sujet. Ces articles sont la source du graphe 6.6. Après plusieurs lectures, ils ont été classés par date de parution. Leur forme indique la nature de leurs auteurs, leur couleur indique la famille des algorithmes de placement

article	noeuds / nodes	liens / edges	source
[12]	20	23	figure 5
[5]	12	42	chapitre 6, réseau abilene, de SNDlib ("SNDlib ¨ 1.0–Survivable Network Design Library,")
[11]	24	37	7.1 Simulation environment, Internet Topology Zoo, BT Europe topology
	110	148	7.1 Simulation environment, Internet Topology Zoo, Interoute topology
[10]	12	15	E. Simulation Setup, Internet2 research network
	23	42	E. Simulation Setup, university data center network
	79	147	E. Simulation Setup, IAS-3967 Rocketfuel topology dataset
[50]	31	72	4.1.1 Experimental Results. Note : 12 overlay nodes + 19 underlay elements
[34]	?	?	VI. Numerical Results, Fat-tree
	?	?	VI. Numerical Results, VL2
	?	?	VI. Numerical Results, B-Cube
	190	260	VI. Numerical Results, inter-data-center network - Cogent, 40 Cogent data center
[21]	16	?	VI. Numerical Result, 16 machines physiques
	40	?	VI. Numerical Result, 40 machines physiques
[24]	12	42	réseau abilene, de SNDlib ("SNDlib ¨ 1.0–Survivable Network Design Library,")
[22]	100	125	Figure 7
[25]	190	260	VI. Performance Evaluation, inter-data-center network - Cogent, https ://cogentco.com/en/network/network-map
[27]	30	?	4.1 Setting of Exeprimental Environment, Inet topology generator
[28]	20-30	?	4.1 Simulation Setup,
[49]	100	?	6. Performance Analysis, nombre de bouées
[6]	?	?	VI. Numerical Results, A. Reference scenarios and benchmarks, "topology of Luxembourg City"
[23]	10	13	IV. Numerical Results, texte et figure 3
	20	31	
	40	67	
[7]	?	?	"real backbone network topology of Internet MCI"
[45]	200	?	VII. Performance Evaluation, A. Simulation Setup
[30]	79	147	V. Performance Evaluation, Rocketfuel Autonomous System topologies
[43]	50	?	V. Numerical Results, B. Simulation Setup, Waxman model
	1344	?	V. Numerical Results, B. Simulation Setup, Fat-Tree
[48]	?	?	VI. Evaluation, A. Simulation Conditions, Fat-Tree pour réseau câblé, antennes RRH (remote radio heads) de l'agence nationale des fréquences (FR)
[20]	344	?	VI. Evaluation, C. Performance Comparison, VL2 Topology
	?	?	VI. Evaluation, B. Parallelism-Aware Placement Algorithm Evaluation, Bcube topology
[39]	24	43	USANet topology, venant de Rocketfuel Topology
[44]	?	?	V. Simulation, A. Simulation Settings, Jellyfish et Fat-Tree
[8]	5-300 + 10-600	?	V. Performance Evaluation, A. Evaluation Setup : utilisation de GT-ITM
[9]	17	26	V. Performance Evaluation, figure 2, inspiré d'un réseau de telco allemand
[38]	?	?	2.3 Hierarchical Component Failures, topologie leaf-spine et fat-tree
[13]	23	37	3. Result and discussion, 3.1 Simulation environment, topologie GEANT
[36]	14	22	VII. Performance Evalution, A. Experimental Setup Description and Assumptions, 14 "stacks de switches" et GNS3 (logiciel d'émulation réseau)

TABLE 6.1 – Recensement des nodes et edges ainsi que les modèles repris dans les articles de la section 6.6.

abordé dans l'article. Les liens montrent les références entre les articles. La méthodologie (section 6.6) décrit ce processus de classification.

La période couverte par ces études va de 2014 à 2023. La majeure partie des articles couvre au moins un algorithme heuristique et la présence de plusieurs couples "ILP+heuristique" est notable. Comme le montre l'état de l'art (section 5.8), cela provient du fait que les algorithmes ILP sont utilisés pour comparer les performances de ces derniers en face des algorithmes heuristiques. On constate également que les articles traitant du machine-learning sont majoritairement situés dans la partie droite du graphique.

Les auteurs d'un seul article se présentent comme issus exclusivement du monde industriel. Les autres proviennent du monde académique ou d'une équipe mixte regroupant académique et industriel.

6.3 Critique sur les ordres de grandeur

Si l'on veut schématiser les réseaux, ce sont des appareils d'utilisateurs, mobile ou non, reliés à des centres de données hébergeant des services par des réseaux gérés par des fournisseurs de télécommunication. Les centres de données sont composés de serveurs, de systèmes de stockage et du matériel de télécommunication. Les serveurs fournissent des capacités de calcul et de mémoire. Le modèle NFV de l'ETSI doit permettre la gestion de ces matériels et des ressources disponibles dans le but de pouvoir les attribuer à une requête pour une SFC, le tout sur du matériel au standard de l'industrie.

6.3.1 Critique des ressources

Des informations sur les instances cloud chez Amazon EC2 sont disponibles sur [53]. On y constatera qu'un grand nombre de types de processeurs sont proposés avec la mémoire associée. Par exemple, on notera que dans la catégorie *usage général*, une grosse dizaine de processeurs sont proposés pour des cas d'utilisation spécifiés. A la lecture de ces informations, il est raisonnable de considérer les centres de données d'Amazon comme n'étant pas homogènes du point de vue matériel. Quant à la tarification à la demande, toujours chez Amazon EC2 ([54]), elle varie de quelques cents à quelques dizaines de cents à l'heure selon certains critères. [45] utilise ces données à des fins de quantifications lors de la vérification de son algorithme. Il est possible de récupérer les informations équivalentes chez Microsoft [55] et Google [56]. On constatera également en se rendant sur ces sites que les centres de données sont également hétérogènes en terme de matériel. L'hétérogénéité du matériel se justifie par la proposition de services optimisés pour la demande des clients : serveurs virtuels, à très haute capacité de mémoire, à forte nécessité de ressource de calcul, orienté intelligence artificielle, ...

Le caractère hétérogène du matériel n'est pas relevé dans la littérature. La mise en commun des ressources physiques dans le pot commun des ressources virtuelles gomme cette différence. Les aspects spécifiques de certains matériels sont pris en compte lors de la location de l'espace auprès des fournisseurs cloud ou sont transparents car la requête fait appel à un service à travers une interface/API. Les informations sur les prix permettent de vérifier les gains ou les économies en terme de coûts. La littérature relève ces gains/économies mais ne les quantifie pas systématiquement, par exemple [6,7,22,40]. Quand il y a quantification, c'est toujours de manière relative entre la situation initiale et la situation optimisée.

6.3.2 Critique des noeuds et topologies

L'état de l'art (chapitre 5) montre que les réseaux s'agrandissent tout en se compliquant. Le nombre de noeuds utilisés dans certains documents ne parait pas très élevé alors que l'on s'attendrait à une augmentation du nombre de ceux-ci dans les simulations. Ces dernières se veulent proches de la réalité ou se basent sur des topologies réelles (section 6.1).

Une première explication serait les dates de parution des articles antérieurs à cette croissance des réseaux. Le graphe 6.1 montre la répartition des articles en fonction du nombre de noeuds par rapport à l'année de publication. On constate que la taille des réseaux évolue peu dans le temps. Seuls deux articles se démarquent de la tendance générale : [43] avec 1344 noeuds et [8] avec 900 noeuds. L'alignement des points sur le graphe 6.1 ne semble pas montrer d'augmentation franche du nombre de noeuds utilisés par les simulations.

Une autre explication serait une preuve de conception ne demandant qu'à être explorée plus en avant. L'article ([12]) est dans ce cas de figure. Mais la taille des réseaux ne décolle pas de façon significative en avançant dans le temps alors que le graphe 6.6 montre l'existence de liens entre eux à travers leurs références.

On peut également mettre en parallèle les chiffres de la table 6.1 avec ceux contenus dans le document [14] publié en 2010 et référencé dans [44], où le nombre de serveurs est très largement supérieur (Figure 6.2). En parallèle, l'article de [57] (Microsoft Research, 2009) parle de centres de données de 100 000 serveurs et [15] (2008) parle de 25 000 hôtes, principalement des switches.

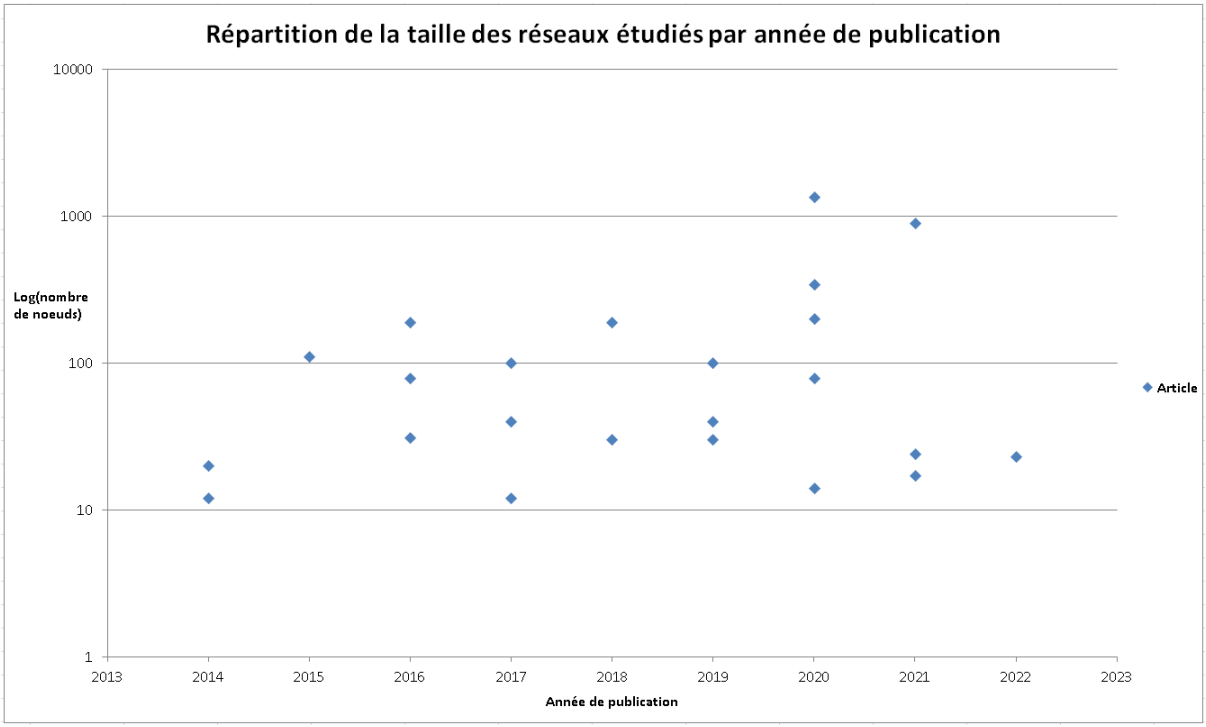

FIGURE 6.1 – Répartition des articles en fonction de leur date de parution et du nombre de noeuds utilisés dans l'article.

Data Center Role	Data Center Name	Location	Age (Years) (Curr Ver/Total)	SNMP	Packet Traces	Topology	Number Devices	Number Servers	Over Subscription
Universities	EDU1	US-Mid	10	✓	✓	✓	22	500	2:1
	EDU2	US-Mid	(7/20)	✓	✓	✓	36	1093	47:1
	EDU3	US-Mid	N/A	✓	✓	✓	1	147	147:1
Private	PRV1	US-Mid	(5/5)	✓	X	✓	96	1088	8:3
	PRV2	US-West	> 5	✓	✓	✓	100	2000	48:10
Commercial	CLD1	US-West	> 5	✓	X	X	562	10K	20:1
	CLD2	US-West	> 5	✓	X	X	763	15K	20:1
	CLD3	US-East	> 5	✓	X	X	612	12K	20:1
	CLD4	S. America	(3/3)	✓	X	X	427	10K	20:1
	CLD5	S. America	(3/3)	✓	X	X	427	10K	20:1

FIGURE 6.2 – Résumé des 10 centres de données étudiés, y compris les dispositifs, les types d'informations collectées et le nombre de serveurs, de [14]

Ici, il faut faire la distinction entre la topologie réseau d'un centre de données et d'une topologie issue d'un réseau de télécommunications. Les topologies telles USANet [39], GEANT [13], IAS-3967 [10], BT Europe [11] et Interoute [11] sont représentatives d'un réseau d'un opérateur de télécommunications. Les noeuds de ces topologies sont des centres de données sur lesquels vont se connecter d'autres opérateurs, comme les opérateurs régionaux ou les fournisseurs d'accès Internet (FAI). En mai 2021, Brain Daigle publie son rapport [58] auprès de la commission US du commerce international où il arrondit le nombre de centres de données autour de 8000, nombre appelé à augmenter. La topologie d'un centre de données se distingue de cette topologie avec une structure plus hiérarchique, comme le montrent [57] et [15] dont le graphe 6.3 est issu. Il convient dès lors de mettre en parallèle le sujet des articles issus de la méthodologie (section 6.6) et la topologie utilisée dans le-dit article pour effectuer une simulation réaliste.

Pourtant l'ordre de grandeur entre les topologies "centre de données" et les chiffres provenant d'autres articles est conséquent. Si on met en parallèle les chiffres du tableau 6.1 et ceux de la figure 6.2, on constate que les simulations se font sur base de centres de données de taille correspondante à celle des types *Universities* et *Private*, le cas des grands centres de données *Commercial* est envisagé dans deux articles ([43] et [8]).

Pour les mêmes raisons que [43] et [8], cela s'explique pour les recherches sur des aspects particuliers, comme la recherche océanique [49] et le réseau de drones [37]. Ces cas d'études n'utilisent pas de réseau de très grande taille.

Une autre explication fait suite à l'analyse des colonnes *Qualité Rédacteurs* et *soutien - financement* du tableau 6.2. Rares sont les grandes entreprises style GAFAM impliquées dans les documents trouvés (6.6), tout au plus retrouve-t-on

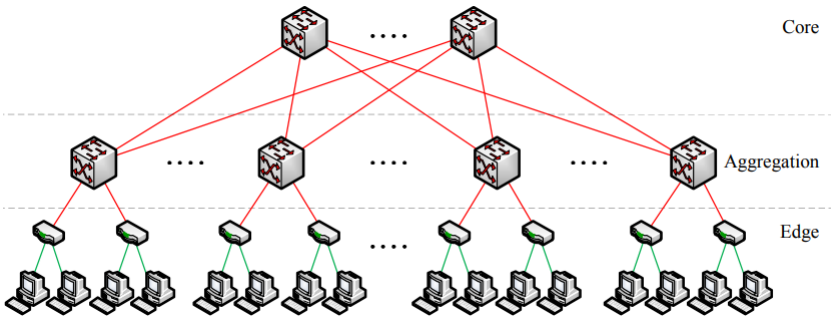

FIGURE 6.3 – Topologie commune d'interconnexion des centres de données, de [15]

des membres d'IBM parmi les chercheurs ou Cisco, Intel et Telecom Italia dans les soutiens de certains articles. Il a fallu chercher dans les références des articles pour trouver le document [57] venant d'un industriel. L'analyse des soutiens des articles tend à montrer que la recherche se fait en université ou dans des centres de recherches subventionnés par des fonds publics. Le cas de la Chine est interpellant car on retrouve dans une part non négligeable des articles faisant référence à des fonds chinois comme soutien à la recherche. Il n'est pas aisé de connaître les éventuelles parties prenantes industrielles à ces fonds.

6.3.3 Critique des requêtes

Le chapitre 5 montre également qu'un des axes de développement des algorithmes est de gérer le nombre de requêtes entrantes. L'article [25] (2018) se base sur le nombre de requêtes HTTP de Wikipedia pour simuler un trafic cohérent avec la réalité. Les auteurs sont récupérés ces informations dans l'article de G. Urdaneta, G. Pierre, and M. van Steen, *"Wikipedia workload analysis for decentralized hosting,"*, paru en 2010 dans *Computer Networks*. Le nombre de requêtes HTTP est de 20,6 milliards en 10 mois, soit une moyenne arrondie de 67 millions de requêtes par jour. Selon un article en ligne sur le site *Le Soir*, [59] donnait en 2018 un chiffre de 3,3 milliards de requêtes par jour et 4 milliards de vidéos visionnées sur Youtube. Certes, les chiffres de [59] sont sujets à caution ; ils devraient être confirmés et affinés mais ils donnent néanmoins une approximation du nombre de requêtes, un rapport de 50 pour 1 en faveur de Google. Les chiffres utilisés en 2018 par [25] datent de 2010. La question du pourquoi prendre des chiffres datant de 8 ans se pose. Une piste de réflexion se trouve dans [30]. Cet article date de 2020 et les auteurs affirment que *"En raison du manque d'ensembles de données NFV publiques, nous synthétisons des topologies de superposition VNF basées sur des topologies Internet réelles."*. Il faut envisager la possibilité que *"En raison du manque d'ensembles de données NFV publiques"* soit également une partie de la réponse. Ce point du manque de données fiables pour permettre l'évaluation correcte des algorithmes est débattu plus loin.

6.4 Critique sur le Machine-Learning

L'état de l'art a abordé l'utilisation du machine-learning. On retrouve cette thématique dans les articles [50], [40], [7], [39], [13], [46] et [35] issus de la recherche documentaire (section 6.6). La connaissance en temps réel des données d'état relatives à l'ensemble des composants des SFC et de l'infrastructure permet l'adaptation en ligne des ressources allouées à ces mêmes composants. Avec des techniques d'apprentissage, le comportement du réseau dans le temps permet de prédire le comportement futur et d'anticiper le placement des composants, idéalement dans des périodes propices au déploiement/déplacement de VNF. A terme, cela permet également d'automatiser la gestion des SFC sur des nouveaux supports technologiques comme la 6G et les réseaux intelligents du futur. L'ensemble des articles repris dans la section 6.6 traitant du machine-learning propose une critique de l'emploi de celui-ci.

Pour reprendre l'idée de la critique sur les ordres de grandeur (section 6.3), si l'on regarde les chiffres du Tableau 6.1, on constatera que pour les articles concernés, le nombre de noeuds/liens utilisés sont petits. [50] met à disposition sur son site web (*https ://knowledgedefinednetworking.org*) des ensembles de données avec des noeuds allant de 14 à 50 en précisant dans sa conclusion la forte dépendance des techniques de machine-learning à la disponibilité d'ensembles de données, normalisées. Toujours dans ce cadre, les expériences menées dans [50] portent sur le routage et les VNF sur du SDN, ce qui explique les topologies proposées. [39] utilise un ensemble de données basé sur USA-Net (*https ://www.researchgate.net/figure/Network-topology-used-in-our-study-a-NSFNET-topology-b-USANET-topology_fig4_254009153*) avec 24 noeuds et 43 liens. Les auteurs placent les VNF dans le cadre des réseaux cellulaires sans fil dans leur introduction. [13] utilise l'ensemble de données GEANT avec 23 noeuds et 37 liens et reprend le réseau défini

par SDN. On retrouve le plan de données et le plan de contrôle auxquels s'interface leur solution. Les topologies sont donc cohérentes avec l'utilisation du machine-learning.

6.4.1 Critique des ensembles de données

L'étude du comportement du réseau est aussi un aspect justifiant l'utilisation du machine-learning. Pour [39] et [13], le comportement est généré par les chercheurs. Cela leur permet de démontrer l'efficacité de leur solution respective. La question se pose de la représentativité du comportement généré par rapport à la réalité. [50] tente de reproduire le comportement en se basant sur les traces d'un campus et le logiciel *tcpreplay* qui reproduit le trafic réseau préalablement capturé. Il y a sélection de 86 caractéristiques de trafic au préalable avant le rejeu des données. [40], [7], [35] et [46] tout au plus citent d'autres ensembles de données dans leurs travaux mais ne les quantifient pas directement.

En réalité, lorsque les auteurs de [50] écrivent leur conclusion, ils attirent l'attention sur les défis de l'utilisation du machine-learning. Ils indiquent dans leur document la nécessité d'adapter le machine-learning pour l'étude des réseaux informatiques. Pour les citer, *"Les graphes sont un exemple notable dans les réseaux pour représenter les topologies, qui déterminent les performances et les caractéristiques d'un réseau. Dans ce contexte, seules des tentatives préliminaires ont été proposées dans la littérature pour créer des algorithmes de machine-learning solides capables de modéliser la topologie des systèmes qui peuvent être représentés par un graphe [19]."* Au moment d'écrire leur article, les auteurs se positionnent au début des travaux sur l'application du machine-learning et rappellent que pour bien fonctionner, les modèles de machine-learning ont besoin d'un ensemble de données d'apprentissage suffisamment représentatif. *"Dans de nombreux cas, les progrès des techniques de ML dépendent fortement de la disponibilité d'ensembles de données normalisés."* [50] et les auteurs précisent que pour certains chercheurs l'utilisation d'ensembles de données de haute qualité priment sur la recherche de nouveaux algorithmes. Or il reste à découvrir, selon eux, ce qu'est un modèle de données d'apprentissage suffisamment représentatif tout en s'interrogeant sur l'existence d'un ensemble d'entraînement représentatif à cause de la variabilité des réseaux.

Les auteurs de l'étude [40] (2019) sont parfaitement explicites sur le problème des ensembles de données d'apprentissage. Dans la section 5.1 de leur document, le dernier paragraphe traitant de cet aspect des ensembles de données commence par : *"Un schéma récurrent commun qui apparaît dans la plupart des articles examinés est que les modèles d'apprentissage automatique proposés n'ont pas été formés et testés à l'aide de grands ensembles de données de haute qualité recueillis auprès d'acteurs industriels puissants dans des environnements de production."* Le reste du paragraphe poursuit sur cette critique, notamment avec la constatation que les ensembles de données utilisés sont synthétiques, trop petits, trop anciens ou non représentatifs de la réalité. Les résultats fournis sont dès lors difficiles à évaluer alors que dans l'introduction, les auteurs soulignent la nécessité d'évaluer les solutions proposées dans un environnement réel. C'est pourquoi les auteurs expriment le besoin d'ensembles de données réalistes et accessibles provenant des grands de l'industrie, expression reprise dans la conclusion. Par ailleurs, les auteurs constatent également la nécessité d'impliquer ces grands de l'industrie dans le processus de définition des exigences et dans l'évaluation des résultats sur du réel. La question de l'implication des industriels est également abordée dans la section 6.3.2.

Début 2023, pour les auteurs de [46], des questions restent en suspens quant à l'utilisation efficace du DRL. Ils signalent dans la section *"Evaluation Challenge"* que la plupart des travaux sont des simulations. L'emploi d'un banc de test réel permettrait des évaluations approfondies pour juger de l'efficacité de la méthode DRL. Cette remarque dans les défis de l'évaluation appuie indirectement sur l'importance d'ensembles de données représentatives. La difficulté à évaluer correctement les travaux sur des simulations en laboratoire incite les auteurs à proposer de passer sur des bancs de tests réels, avec des données et comportements réels. La remarque de [40] est confirmée 4 ans plus tard mais n'est cependant pas le seul article de 2023 à aller dans ce sens. Les auteurs de [52] préconisent dans le point *"7.3 Usable data"* de générer des ensembles des données similaires à celles d'ensembles de données réelles et de passer par des bancs d'essais réels. Ceci rejoint un des points de la conclusion de [50] qui préconise l'utilisation d'ensembles de données normalisées.

Malgré le besoin exprimé à plusieurs reprises d'avoir des ensembles de données de haute qualité pour pouvoir valider les recherches, il est un facteur qui rend la récupération de ces données pour générer les ensembles de données attendus : la confidentialité. La probabilité est quasi certaine que pendant la capture des données, il se trouve des données à caractère confidentiel. C'est un problème déjà relevé par [40] qui propose de n'utiliser qu'un sous-ensemble de données ainsi que leurs modèles de génération de charge. [35] rappelle également que les informations collectées doivent être gérées intelligemment pour assurer, entre autre, la confidentialité. Les auteurs de [46] pointent également la problématique de collecter certains états qui sont également liés à la confidentialité et, de ce fait, l'exploitation des données se fera sur un sous-ensemble, avec des conséquences comme une partie du comportement non mesuré. Les auteurs de [52] rappellent à très juste titre que les données qui transitent sur le réseau sont sujettes au GDPR. Outre l'aspect administratif de la gestion des données imposée par cette réglementation, cela devient une obligation de prendre en compte l'aspect confidentiel des

données collectées et l'application de solutions automatisées d'anonymisation ou de filtrage peut potentiellement ralentir la collecte de ces données. Selon [52], il reste de nombreuses craintes dans le monde de la recherche, dont la fuite d'informations sensibles de l'ensemble original vers l'ensemble de données synthétiques.

6.4.2 Critique des limites du machine-learning

La première critique du machine-learning vient d'être formulée et concerne la limite des ensembles de données utilisables pour permettre de valider les travaux.

La seconde vient de l'importance du changement de paradigme que constitue l'utilisation du machine-learning. Dans leur article de juillet 2023, les auteurs de [52] posent dans leur introduction la question de l'utilité du machine-learning ou l'IA dans la gestion du réseau et constatent également que *"[d]e toute évidence, de nombreuses propositions ont été faites pour rendre les réseaux plus intelligents grâce à l'apprentissage automatique, mais elles n'ont pas été adoptées à grande échelle pour l'exploitation de réseaux réels, car elles modifient les paradigmes en faveur de méthodes stochastiques."* Les paradigmes évoqués sont les méthodes ILP et heuristiques utilisées pour le placement de VNF. Ces méthodes ont des résultats compréhensibles par les humains et, selon [35], l'automatisation se fait par script, avec les limitations qu'imposent les conditions d'utilisation de ces scripts. L'IA et le machine-learning s'orientent vers plus d'automatisation autonome, ne nécessitant pas d'intervention humaine. Cela ne peut se faire qu'avec l'acceptation et la confiance des gens qui vont employer ces solutions sur les infrastructures critiques que sont les réseaux. [52].

Une autre critique vient de l'utilisation du machine-learning. C'est une technologie appliquée aux réseaux relativement récemment. Á ce stade, elle n'a pas été adoptée à grande échelle pour l'exploitation des réseaux [52]. Les critiques précédentes comptent pour partie dans la réponse. D'autres points peuvent être ajoutés. Ce même article indique aussi une autre limitation de l'IA, à savoir la difficulté à s'adapter au volume et à l'hétérogénéité des données récupérables sur un réseau. On peut ajouter qu'en cas d'itération de l'algorithme de DRL, le temps de convergence vers la solution peut être long ([46]) quand ce n'est pas un temps de calcul conséquent de la part de la méthode choisie ([40]). L'emploi de matériel hautement spécialisé est parfois nécessaire pour l'utilisation de certaines techniques comme le Deep Neural Network (DNN). A l'avenir, [35] annonce que du matériel dédicacé à l'exécution d'algorithmes de machine-learning sera utilisé pour une exploitation distribuée.

6.5 Discussion

L'état de l'art et ses critiques ont retracé les grandes étapes chronologiques du placement des VNF, allant de l'énoncé de concepts nécessaires que sont la SFC, l'architecture NFV à l'utilisation de ces concepts dans les futurs réseaux informatiques. Les grandes familles algorithmiques ont été abordées ainsi que les différents critères qui influencent la mise en place des VNF dans des infrastructures NFV et les critiques ont mis en avant des problèmes de topologies, de représentation du comportement du réseau et d'emploi de certaines techniques.

Pour le problème de topologies, il faut séparer les topologies des opérateurs de télécommunications des opérateurs de centre de données. Il faut peu de temps de recherche pour découvrir les topologies des opérateurs et certains articles les reprennent explicitement dans leurs références. On peut raisonnablement conclure que ces topologies sont réalistes. Pour les centres de données, la vision de [15], fig 1 (ou figure 6.3) est également reprise dans [57], fig 1 (de Cisco, 2004) et [14], fig 1. On peut considérer que cette vision est cohérente avec les centres de données actuels. En l'état de mes connaissances, il me semble que cette vision est toujours d'application. Pour un opérateur de centre de données, une fois la certitude acquise que les réponses aux demandes entrantes sont envoyées, le transport vers la destination n'est plus de sa responsabilité ; ses connexions avec le reste du monde est un service qu'il paye. Si cet opérateur gère son centre de données avec une architecture NFV, ce qui sort est hors de son domaine. Son algorithme de placement de VNF se limitera à son centre de données. Le raisonnement est identique pour chaque opérateur sur le trajet. Au final, une requête initiale va générer un certain nombre de requêtes de création de SFC sur son chemin, en fonction du nombre d'opérateurs que la requête va traverser.

Une autre réflexion concernant les centres de données est l'utilisation de ces derniers par des clients tiers. Il est envisageable qu'une partie du centre de données soit louée par un client et qu'à ce titre, elle soit isolée du reste du centre de données jusqu'à un niveau proche de la connexion de sortie. Cela a pour effet de fractionner virtuellement le centre de données. Il a été également constaté dans la section 6.3.1 que les grands opérateurs proposaient des serveurs à utilisation spécialisées. La question de la gestion de cette diversité de matériel et de l'impact de cette diversification sur l'orchestration des VNF serait intéressante à investiguer. Même si les centres de données sont de taille conséquente, l'ordre de

grandeur utilisé par [43] et [8] est cohérente avec l'ordre de grandeur des centres de données *commercial* du graphique 6.2 si on admet que les serveurs hébergés dans un centre de données *commercial* ne forment pas un ensemble monolithique. La colonne *over subscription* laisse déjà entrevoir une répartition. L'idéal serait d'avoir un retour des opérateurs de ces grands centres de données mais il peut relever de l'intérêt économique de ces derniers de ne pas communiquer à ce sujet. Une organisation particulière et efficace de leurs centres de données peut leur apporter un avantage concurrentiel sur les autres acteurs. La comparaison de l'orchestration des grands centres de données avec des centres de données plus petits est une piste pour de futures recherches à mener avec les partenaires industriels. Cela permettrait de comparer les comportements des orchestrations par rapport au changement d'échelle.

Dans l'état de l'art, des bémols ont été repérés dans la littérature. On peut douter de l'utilisation d'un système complet NFV orchestré par un MANO tel que décrit par l'ETSI ([9]). C'est un point qui serait tranché indirectement par la question posée sur la comparaison des comportements des orchestrations. Pour le second bémol, la performance des VNF versus middleboxes physiques, ce qui est clairement énoncé dans [10], ce sont des performances pour les VNF proches des performances du matériel. Dans [34], des solutions SDN et NFV sont utilisées depuis 2016 par les opérateurs de télécommunications que sont AT&T et Verizon. Avec des grands opérateurs utilisant cette technologie depuis 2016 et la recherche continuant avec, en perspective, l'utilisation sur les nouveaux réseaux informatiques (6G et réseaux intelligents), notamment les évolutions constatées dans le groupe de travail de l'ETSI, on peut en déduire que cette approche VNF rencontre une partie sinon tous ses objectifs. Si on s'intéresse un peu aux propositions faites par les fournisseurs de solution VNF, la question de la comparaison de performances entre VNF et middleboxes physiques pourrait faire une étude intéressante. En admettant que l'assertion de [12] soit toujours pertinente, comment et jusqu'où la performance moindre d'une VNF se justifie-t-elle ? Car l'adoption de la technologie NFV a aussi certains inconvénients.

La virtualisation des middleboxes a apporté de la souplesse pour la création et la maintenance des SFC. Les techniques de création automatique de machines virtuelles, préconfigurées ont bien évolué. Cependant, lorsqu'on cherche à créer une plateforme NFV complète, on doit rapidement gérer plusieurs solutions logicielles pour gérer l'un ou l'autre composant. L'assertion initiale parlait de personnel qualifié pour des middleboxes propriétaires, ce qui avait un impact sur les coûts de maintenance. Le problème soulevé à l'époque est toujours existant. Certes, la virtualisation a permis une maintenance plus aisée, supprimant de fait les problèmes matériels, ce qui est un gain important tant la possibilité de faire un *"snapshot"* (une sauvegarde de l'état d'une machine virtuelle à un moment donné) sur une machine virtuelle sécurise les manipulations sur la-dite machine. Mais les opérations sur certaines machines demandent toujours des connaissances approfondies sur leur utilisation. De plus, un problème sur l'infrastructure physique ou sur les logiciels de virtualisation/de gestion demande également des compétences spécifiques sur du matériel aux standards de l'industrie mais développé par des firmes (comme HP et son infrastructure de centre de données) ou des logiciels spécifiquement développés pour l'infrastructure NFVI. Ce phénomène de glissement des spécialisations nécessaires au bon fonctionnement est encore accentué par la venue du machine-learning.

Le machine-learning est incontestablement un sujet chaud du moment. ChatGPT et les images construites par AI pour ne citer que ces deux exemples ont porté le sujet de l'intelligence artificielle vers le grand public. Or, c'est un sujet incroyablement complexe et difficile à appréhender sans un minimum de connaissance théorique. L'état de l'art explique grossièrement les buts de l'introduction du machine-learning. Avec la capacité d'apprendre le comportement du réseau, la partie machine-learning peut produire un modèle basé sur la statistique et, en fonction de certaines règles, modifier les ressources allouées aux VNF en cas de creux ou anticiper le placement de ces dernières afin de prévenir un pic prévisible du trafic réseau, de manière automatisée et "en ligne". Pourtant, dans la littérature citée jusqu'à présent sur ce sujet, les auteurs signalent que le comportement du réseau peut être très variable. La réaction du système en cas de pic statistiquement imprévisible, comme une attaque par flood, ou de creux est une réalité à prendre en compte dans le développement. Il y a aussi le scénario d'éventuels enchaînements de ces deux phases. L'apparition d'un type de ces incidents a une probabilité non négligeable de se produire. La question de la pertinence d'inclure ces scénarios revient à faire la balance entre un système plus robuste mais impliquant plus de ressources et un système moins robuste mais plus rentable, moins énergivore, moins consommateur de ressources, supportant plus de trafic. Dans le même ordre d'idée, il est fait référence dans les critiques des limites du machine-learning à certaines techniques nécessitant du matériel hautement spécialisé. La pondération des gains obtenus en utilisant ces matériels spécialisés devrait idéalement compenser la consommation énergétique de ceux-ci.

Un autre sujet machine-learning est la critique faite dans plusieurs articles concernant le manque d'ensembles de données réalistes et de haute qualité pour permettre à la recherche de valider leurs travaux. La critique des ensembles de données (section 6.4.1) rappelle également les problèmes de confidentialité qui requièrent de créer un sous-ensemble expurgé de ces données problématiques. En parallèle, on peut également rappeler que la technique du découpage en tranches de réseaux (*slicing*) permet de regrouper le trafic réseau de certains types d'applications et d'isoler leur trafic des autres tranches. Cela revient à catégoriser une partie du trafic, dans le sens de [1]. Dans cette logique, il devrait être possible de

définir des SFC types pour des applications spécifiques comme le streaming 4K, par exemple. Avec des pré-conditions généralistes, sur un modèle topologique physique défini, la simulation dans ce cadre strict permettrait de comparer les résultats entre différentes solutions. L'IETF semble être l'endroit le plus pertinent pour mener cette tâche à bien car on y retrouve les grands acteurs en matière de trafic réseau. Ces derniers pourraient estimer le comportement réseau d'un type d'application en particulier, comme le trafic des plateformes de streaming.

Les documents trouvés dans la section 6.6 font principalement suite à des projets menés par des institutions publiques, avec relativement peu de soutien du monde de l'industrie. La section 6.3.2 indique également que peu d'industriels sont mentionnés parmi les auteurs. On pourrait s'en étonner alors que la présence de certains soutiens est à signaler, comme les secteurs de la Défense (UK et US, DARPA) ou des secteurs travaillant dans des environnements spécifiques, comme l'océanographie ou les drones. Pour atténuer l'étonnement, plusieurs remarques sont à émettre. Les grands groupes industriels intéressés sont présents dans le groupe de travail de l'ETSI ou dans les groupes de l'IETF. A défaut de soutenir directement des études, ils le font à travers les organisations précitées. Certains grands groupes proposent des solutions commerciales ou ont implémenté ces solutions en leur sein. Ainsi, Microsoft propose des VNF depuis sa version 2016 de Windows Server ([60]) déployables sur ses services de virtualisation. Il est précisé pour certaines VNF qu'elles sont similaires à la version exploitée sur Azure.

Enfin, un dernier point qui n'a pas été développé dans l'état de l'art et dans la critique mais qu'il me semble intéressant d'aborder est l'impact des centres de données. Comme dit dans le rapport [58], il y avait à peu près 8000 centres de données. En recherchant cette information, j'ai trouvé d'autres rapports ([61] et [16]) sur le poids économique des centres de données. Outre la croissance du nombre des centres de données, la puissance électrique totale consommée par tous ces centres est estimée à 20% de la production mondiale vers 2025 [61] alors que l'efficacité énergétique stagne depuis une dizaine d'année (figure 6.4). Les propriétaires de ces centres de données installent dorénavant des moyens de production et de stockage [16]. La recherche et développement pour améliorer le refroidissement et l'amélioration énergétique des centres de données devrait être intensifiée [16]. Il y a là quelques questions de recherche à se poser. L'utilisation du machine-learning est-elle plus consommatrice d'énergie que les autres méthodes pour fournir une solution ? L'utilisation du machine-learning permet-elle de diminuer significativement la consommation du reste des composants réseaux et NFVI ? La recherche de solution ne devrait-elle pas prendre davantage en compte le critère de gains énergétiques ? Enfin, il y a les questions plus éthiques sur le fait de rajouter des centres de données grands consommateurs d'énergie alors que nous vivons en ce moment des changements importants, que ce soit climatique ou énergétique auxquels les sociétés humaines doivent apporter une réponse.

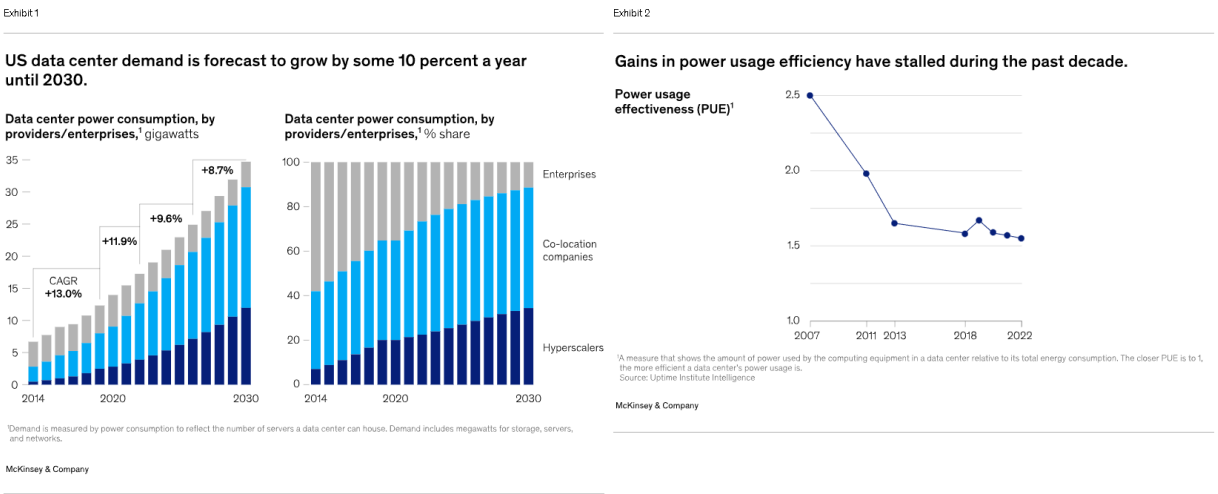

FIGURE 6.4 – Graphiques extraits de [16]

6.6 Méthodologie

Ce travail s'appuie sur des documents trouvés initialement par une recherche par mot clé, à savoir **"vnf placement", algorithm, verification**. Plusieurs moteurs de recherche ont permis de relever un nombre raisonnable d'articles pour constituer une bibliographie de départ. Lors de la première itération de la recherche a donné le résultat suivant :

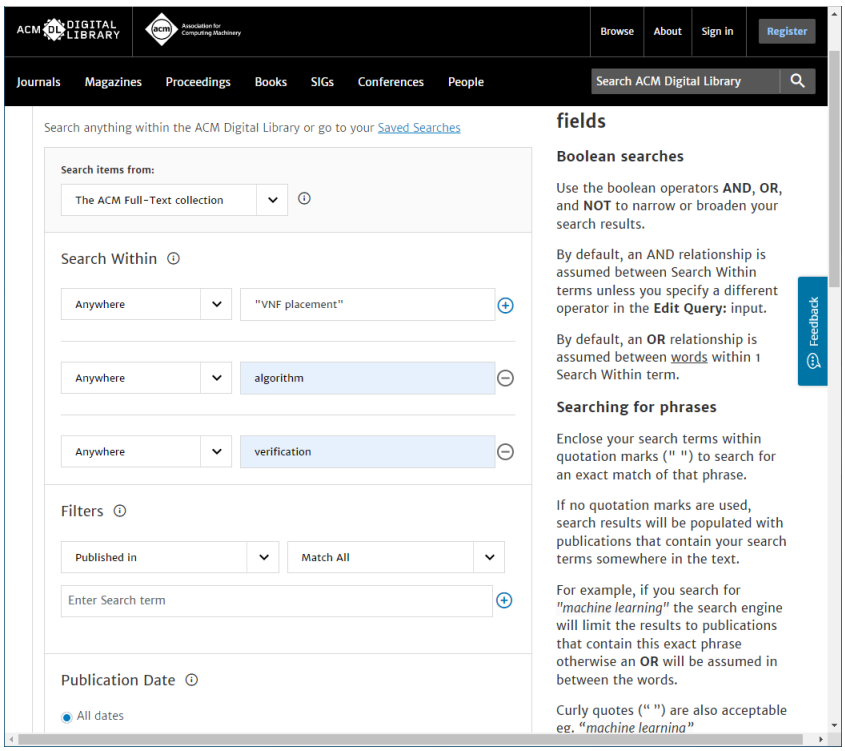

FIGURE 6.5 – Capture d'écran de la recherche sur https ://dl.acm.org/search/advanced

Moteur de recherche	nombre d'articles
IEEEXplore	0
arXiv	0
ACM	24
IETF	0
Google Scholar	257

Lors d'une itération suivante, en novembre 2022, des articles supplémentaires ont été ajoutés à la liste, passant ainsi à 26. Lors de la rédaction, une nouvelle itération de la recherche a été refaite sur le site de l'ACM et le décompte des articles est passé à 30. Ces 4 articles n'ont pas été pris en compte dans ce travail.

Lors de cette phase de recherche documentaire et tout au long de ce travail, mon promoteur a fourni plusieurs articles supplémentaires pour compléter la bibliographie : *Deep Reinforcement Learning Approaches to Network Slice Scaling and Placement : A Survey* [46], *Service Chain Placement Optimization in 5G FANET-Based Network Edge* [37] et *Resource Abstractions in NFV Management and Orchestration : Experimental Evaluation* [36] avec un paragraphe traitant plus particulièrement du sujet de ce document. Ce paragraphe citant d'autres articles [10], [24], [11], [9, 23, 39] , ces derniers ont été ajoutés à la bibliographie.

Après cette collecte de documentations et une lecture des *abstracts*, introductions et conclusions pour vérifier la pertinence supposée du contenu, un graphe représentant les citations entre les documents est créé. On peut constater que toutes les branches du graphe ne sont pas reliées entre elles. Une recherche dans les références des documents a été effectuée pour découvrir un éventuel article commun. Deux textes, *Specifying and Placing Chains of Virtual Network Functions et VNF-P : A Model for Efficient Placement of Virtualized Network Functions* [5, 12], ont ainsi été trouvés et ajoutés à la bibliographie. Au fil de ses lectures, mon promoteur m'a transmis les articles *Network Intelligence for NFV Scaling in Closed-Loop Architectures* et *Research Challenges in Coupling Artificial Intelligence and Network Management* [52] qui pouvaient être en lien avec ce travail.

Le graphe 6.6 est monté de la manière suivante :

1. Chaque noeud du graphe représente un article et se voit attribuer un numéro.
2. Chaque noeud a une taille fonction du nombre de citations. Le nombre de citations vient de Google Scholar. Cet outil de recherche a été choisi car il retournait une réponse pour chaque article. La recherche sur le nombre de citations a été effectuée en avril 2023.

3. Les noeuds sont disposés sur une ligne du temps et la date de parution est ajoutée pour faciliter la lisibilité. La date de parution est la date renseignée sur le site depuis lequel le document a été téléchargé.
4. Une couleur est attribuée à chaque noeud après lecture de l'article et sa classification par type d'algorithme. Certains articles traitent de plusieurs familles d'algorithmes. Il est à noter que certains articles utilisent plusieurs types d'algorithmes, généralement dans le but de comparaison de résultat, comme dans [11]. 4 familles ont été mises en évidence : les algorithmes (M)ILP (rouge sur le graphe 6.6), heuristiques (jaune), machine learning (bleu) et autres (blanc). Des catégories mixtes (orange, verte) ont aussi été créés. La catégorie "autre" reprend ce qui ne rentre pas dans les 3 première catégories d'algorithmes.
5. Le lien dirigé reliant 2 noeuds signifie que la source est une référence présente dans l'article cible. La couleur du lien sert à distinguer les liens entre eux.

Par la suite, tout au long de l'étude des documents, de nouveaux articles sont venus s'ajouter. Certains proviennent des références de certains articles listés précédemment. Les informations, utiles au sujet de ce travail et contenues dans ces articles référencés, justifient leurs citations et l'ajout à la bibliographie.

Enfin, des informations disponibles sur certains sites Internet et autres rapports publiés en ligne m'ont permis de justifier certains arguments ou détails. Ils font également partie des références.

Date Publication	Référence	Citations Google Scholar	Qualité Rédacteurs	Soutien Financement
novembre 2014	[12]	544	Académique	Institute for the Promotion of Innovation by Science and Technology in Flanders
décembre 2014	[5]	610	Académique	International Graduate School "Dynamic Intelligent Systems"
octobre 2015	[11]	173	Académique Industriel	FP5 UNIFY projet (EU), BOF/GOA project (UGent)
mai 2016	[10]	376	Académique	National Sciences and Engineering Research Council of Canada, Smart Applications on Virtual Infrastructure (SAVI) project, University of Waterloo
juin 2016	[50]	379	Académique	Spanish Ministry of Economy, Industry and Competitiveness, EU FEDER, Catalan Government
juillet 2016	[34]	359	Académique	non précisé
janvier 2017	[24]	23	Académique	SONATA project (Horizon 2020/5G-PPP programs, EU Commission), German Research Foundation (DFG)
janvier 2017	[21]	49	Académique	Quatar National Research Fund
mars 2017	[22]	321	Académique	TELECOM ITALIA
février 2018	[25]	106	Académique Industriel	non précisé
mai 2018	[26]	9	Académique	National Science and Technology Major Project of China, Fundamental Research Funds for the Central Universities, National Natural Science Foundation of China, Collaborative Innovation Center of Novel Software Technology and Industrialization, Jiangsu Innovation and Entrepreneurship (Shuangchuang) Program
novembre 2018	[27]	9	Académique	National Key RD Program of China, EU-China study on IoT and 5G(EXCITING), Research Fund of Ministry of Education-China Mobile
janvier 2019	[47]	113	Académique	H2020 projects 5G-TRANSFORMER, (EU Commission)
juin 2019	[41]	182	Académique Industriel	U.S. Army Research Laboratory, U.K. Ministry of Defence
août 2019	[28]	12	Académique	U.S. NAtional Science Foundation
août 2019	[29]	9	Académique	National Key RD Program of China, National Natural Science Foundation of China,

Date Publication	Référence	Citations Google Scholar	Qualité Rédacteurs	Soutien Financement
				Natural Science Foundation of Jiangsu Province, Fundamental Research Funds for the Central Universities
septembre 2019	[40]	129	Académique	EU Horizon 2020 program, Spanish Ministry of Science, Innovation and Universities, the Comunidad de Madrid
septembre 2019	[49]	0	Académique	Fundamental Research Funds for the Central Universities, National Natural Science Foundation of China, Changzhou Sci. and Tech. Program, Global Infrastructure Program (National Research Foundation of Korea)
octobre 2019	[6]	56	Académique	5GROWTH project (EU Commission), Israel Science Foundation, Neptune Consortium (Israeli Ministry of Economy and Industry)
décembre 2019	[23]	36	Académique Industriel	National Natural Science Foundation of China, Zhejiang Natural Science Foundation, Scientific Research Foundation of Ningbo University, K. C. Wong Magna Fund (Ningbo university)
mars 2020	[45]	24	Académique	Research Grants Council (RGC) of Hong Kong
mars 2020	[7]	29	Académique	National Natural Science Foundation of China, ECNU XingFuZhiHua Program, National Key Research and Development Program of China
avril 2020	[30]	16	Académique Industriel	U.S. Army Research Laboratory, U.K. Ministry of Defence
septembre 2020	[48]	15	Académique Industriel	non précisé
novembre 2020	[51]	8	Académique	Conseil national de développement scientifique et technologique (Brésil), US National Science Foundation, CONIX Research Center (US DARPA), USC's Annenberg Fellowship, Cisco, Intel
décembre 2020	[43]	7	Académique	non précisé
décembre 2020	[20]	13	Académique	Natural Science Foundation of China
janvier 2021	[39]	16	Académique	Nature and Science Foundation of China, National Key Research and Development Program of China
avril 2021	[44]	9	Académique	non précisé

Date Publication	Référence	Citations Google Scholar	Qualité Rédacteurs	Soutien Financement
			Industriel	
mai 2021	[8]	30	Académique	National Key Research and Development Plan, NSFC-Deutsche Forschungsgemeinschaft (DFG), Fundamental Research Funds for the Central Universities National Program for Support of Top-Notch Young Professionals
mai 2021	[42]	8	Académique	EU Celtic Plus/Vinnova project, Health5G - Future eHealth powered by 5G
septembre 2021	[9]	6	Académique	Ministère italien de l'éducation, de l'université et de la recherche, Project 5GROWTH
décembre 2021	[38]	3	Académique	non précisé
juin 2022	[13]	0	Académique	National Natural Science Foundation of China, National Key Research and Development Program of China
juillet 2022	[37]	6	Académique	PIACERI project OMNIA, PRIN Project Liquid-Edge, POR S6 Project
octobre 2022	[36]	0	Académique	Spanish MINECO, Spanish RELAMPAGO (MCIN/AEI)
novembre 2022	[35]	3	Industriel	non précisé
janvier 2023	[46]	0	Académique	non précisé

TABLE 6.2: Tableau reprenant la date de parution, le numéro dans le graphe 6.6, son numéro dans le chapitre 8, le nombre de citations de l'article, la qualité des chercheurs et les soutiens

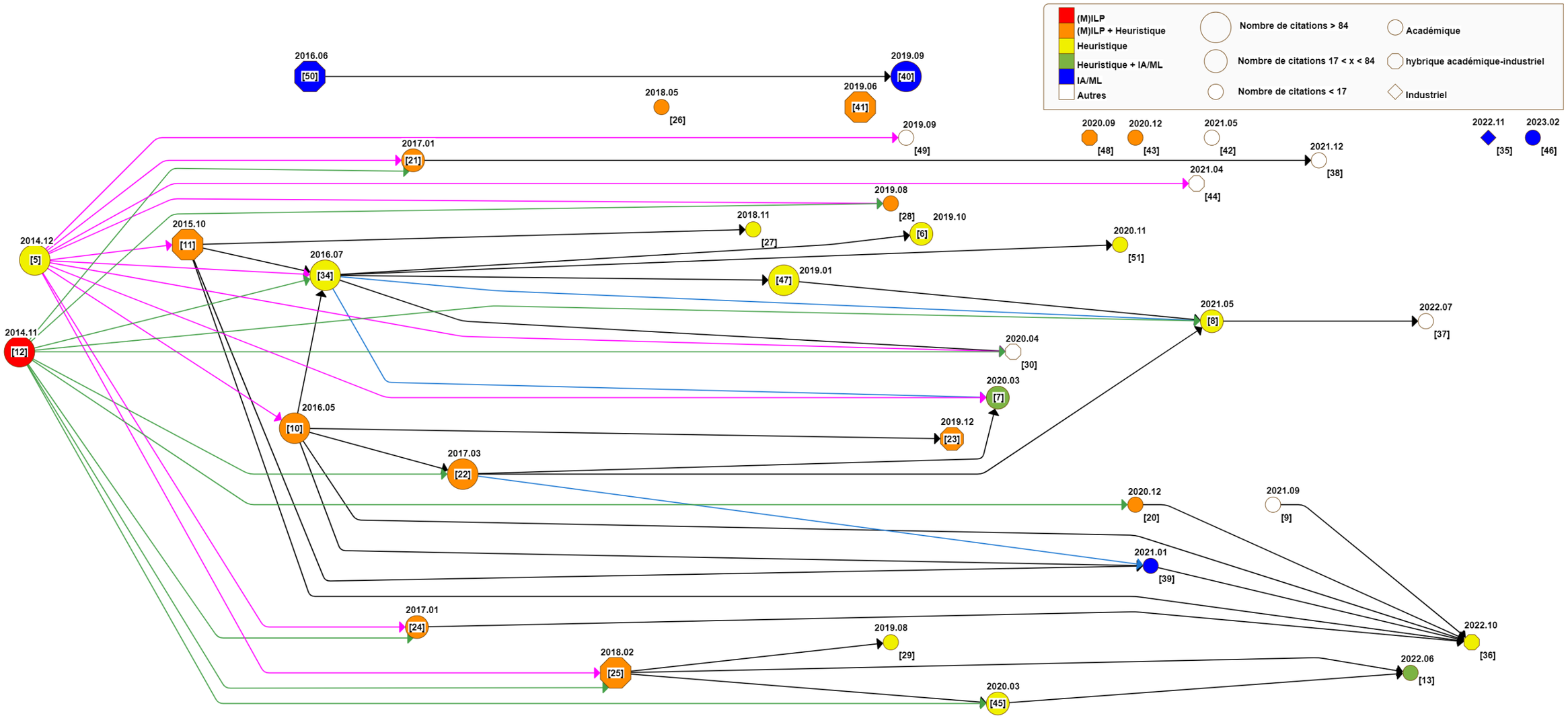

FIGURE 6.6 – Graphe des articles classés par date de parution.

Chapitre 7

Conclusion

A partir d'une recherche sur base de mots clés, une liste de documents a été établie à laquelle s'est ajoutée des documents reçus de mon promoteur. Cet ensemble est la source du graphe 6.6. Une première version de ce graphique montrait les références entre articles ainsi qu'une première classification des méthodes décrites dans les articles pour le placement des VNF. Comme il y avait plusieurs groupes distincts d'articles, une recherche entre des références communes a été effectuée. D'autres articles plus anciens ont été mis en évidence et ajoutés au graphique. Par la suite s'est construite une réflexion autour des mots clés et du contenu des articles. Une nouvelle itération du graphique agrémentée du nombre de citations pour repérer des articles les plus cités est proposée dans ce travail.

L'étude des documents et le graphique 6.6 montrent l'évolution des algorithmes de placement des VNF dans le temps. D'algorithmes ILP résolvant le problème de placement, les chercheurs ont ensuite dirigé leurs recherches sur des algorithmes heuristiques. Cette famille d'algorithmes répondait à la problématique du passage à l'échelle. Les algorithmes ILP étaient trop lents voire incapables de fournir une réponse dans un temps acceptable par les opérateurs. Mais les solutions heuristiques permettent aussi de rencontrer d'autres critères que le graphe 6.6 ne capture pas.

L'avantage des algorithmes heuristiques est de pouvoir mettre l'accent sur un ensemble de critères à privilégier lors de la confection de ces algorithmes. L'état de l'art montre le caractère parfois antagoniste entre deux critères à prendre en compte lors de la création de l'algorithme. Par exemple, déployer un nombre minimal de machines virtuelles et la nécessité d'avoir des machines virtuelles en backup en cas de panne. Le chapitre 5 met également en lumière le lien entre recherche de performance et recherche de rentabilité, l'article [7] en a fait son objet de discussion.

Ce que mettent aussi en avant les documents sont les nouveautés technologiques auxquelles le placement des VNF doit s'adapter. "Edge-computing", "cloud-computing", "slicing", "IoT", 5G et réseaux du futur apportent de nouvelles caractéristiques techniques aux réseaux qui doivent être intégrés au calcul du placement et au maintien des services. Ces spécifications de haute disponibilité et de performance demandent des adaptations automatisées et en ligne du placement. Le nombre d'appareils utilisant des services en ligne étant en croissance, le volume des données transférées est en augmentation ainsi que sa fréquence. La prédiction du comportement du réseau est donc un enjeu à prendre également. Des chercheurs ont donc étudié l'emploi de l'intelligence artificielle et le machine-learning pour résoudre ce problème complexe.

La mise en place de cette technologie est l'objet de recherche. Les critiques du machine-learning (section 6.4) sont abordées concernant les ensembles de données et les difficultés de valider et de comparer les résultats provenant de la recherche. La mise en place de cette technologie ne pose pas seulement ce problème d'ensemble de données et de modélisation du comportement du réseau. La partie "Discussion" met en avant plusieurs aspects de l'apport du machine-learning : adaptation des industries au niveau des compétences nécessaires pour déployer et employer cette solution, la rentabilité de certaines orientations, comme le DNN par rapport à d'autres. La discussion (section 6.5) évoque la possibilité de définir des scénarios industriels génériques utilisables par la recherche à travers l'IETF.

Une dernière considération au niveau de la partie discussion porte sur le sujet des centres de données, sur leur poids économique et énergétique. La technologie du machine-learning nécessite du matériel spécifique. Dans le contexte énergétique et climatique, les questions de consommation électrique nécessaires sont peu abordées dans la littérature. Avec les chiffres donnés dans le paragraphe traitant de ce sujet, le but est d'interpeller les lecteurs et de donner quelques éléments pour amorcer une réflexion.

Un dernier point pour terminer la conclusion de ce travail concerne l'état de l'art. Ce dernier a été structuré pour tenter d'expliquer pas à pas les différents aspects du problème de placement des VNF évoqués dans la littérature trouvée lors

de la recherche documentaire. La volonté première est de donner des clés de compréhension à celles et ceux qui, non familiers du sujet, puissent y entrer. C'est un des buts poursuivi lors de la rédaction.

Chapitre 8

Références

Bibliographie

[1] Joel M. Halpern and Carlos Pignataro. Service Function Chaining (SFC) Architecture. RFC 7665, October 2015.

[2] Carlos J. Bernardos, Akbar Rahman, Juan-Carlos Zúñiga, Luis M. Contreras, Pedro Andres Aranda, and Pierre Lynch. Network Virtualization Research Challenges. RFC 8568, April 2019.

[3] ETSI - site web. Schéma nfv et travaux associés. `https://www.etsi.org/technologies/nfv, lienversNFVarchitecture`, last accessed : 27.07.2023).

[4] Deval Bhamare, Raj Jain, Mohammed Samaka, and Aiman Erbad. A survey on service function chaining. *J. Netw. Comput. Appl.*, 75(C) :138–155, nov 2016.

[5] Sevil Mehraghdam, Matthias Keller, and Holger Karl. Specifying and placing chains of virtual network functions. In *2014 IEEE 3rd International Conference on Cloud Networking (CloudNet)*, pages 7–13, 2014.

[6] Francesco Malandrino, Carla Fabiana Chiasserini, Gil Einziger, and Gabriel Scalosub. Reducing service deployment cost through vnf sharing. *IEEE/ACM Transactions on Networking*, 27(6) :2363–2376, 2019.

[7] Peijin Cong, Guo Xu, Tongquan Wei, and Keqin Li. A survey of profit optimization techniques for cloud providers. *ACM Comput. Surv.*, 53(2), mar 2020.

[8] Qixia Zhang, Fangming Liu, and Chaobing Zeng. Online adaptive interference-aware vnf deployment and migration for 5g network slice. *IEEE/ACM Trans. Netw.*, 29(5) :2115–2128, may 2021.

[9] M. Gharbaoui, B. Martini, G. Cecchetti, and P. Castoldi. Resource orchestration strategies with retrials for latency-sensitive network slicing over distributed telco clouds. *IEEE Access*, 9 :132801–132817, 2021.

[10] Faizul Bari, Shihabur Rahman Chowdhury, Reaz Ahmed, Raouf Boutaba, and Otto Carlos Muniz Bandeira Duarte. Orchestrating virtualized network functions. *IEEE Transactions on Network and Service Management*, 13(4) :725–739, 2016.

[11] Sahel Sahhaf, Wouter Tavernier, Matthias Rost, Stefan Schmid, Didier Colle, Mario Pickavet, and Piet Demeester. Network service chaining with optimized network function embedding supporting service decompositions. *Computer Networks*, 93 :492–505, 2015. Cloud Networking and Communications II.

[12] Moens, Hendrik and De Turck, Filip. VNF-P : a model for efficient placement of virtualized network functions. In *2014 10TH INTERNATIONAL CONFERENCE ON NETWORK AND SERVICE MANAGEMENT (CNSM)*, pages 418–423, 2014.

[13] Di Zhu, Jinchuan Pei, and Yunchen Zhu. A vnf capacity adjustment method based on deep neural network. In *Proceedings of the Asia Conference on Electrical, Power and Computer Engineering*, EPCE '22, New York, NY, USA, 2022. Association for Computing Machinery.

[14] Theophilus Benson, Aditya Akella, and David A. Maltz. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, page 267–280, New York, NY, USA, 2010. Association for Computing Machinery.

[15] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture. *SIGCOMM Comput. Commun. Rev.*, 38(4) :63â€“74, aug 2008.

[16] Srini Bangalore, Bhargs Srivathsan, Arjita Bhan, Andrea Del Miglio, Pankaj Sachdeva, Vijay Sarma, Raman Sharma. Investing in the rising data center economy. `https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy#/`, last accessed : 03.08.2023).

[17] GICAT. La dga a retenu thales/sopra steria et atos pour le traitement massif des données du ministère des armées. `https://www.gicat.com/la-dga-a-retenu-thales-sopra-steria-et-atos-pour-le-traitement-massif-des-donnees-du-\\ministere-des-armees`, last accessed : 11.05.2023).

[18] Ministère des armées (France). Scaf – un chantier en cours, une ambition intacte. `https://www.defense.gouv.fr/actualites/scaf-chantier-cours-ambition-intacte`, lastaccessed : 11.05.2023).

[19] ETSI. Why do we need 5g ? `https://www.etsi.org/technologies/5G`, last accessed : 11.05.2023).

[20] Sihao Xie, Junte Ma, and Jin Zhao. Flexchain : Bridging parallelism and placement for service function chains. *IEEE Transactions on Network and Service Management*, 18(1) :195–208, 2021.

[21] Long Qu, Chadi Assi, Khaled Shaban, and Maurice Khabbaz. Reliability-aware service provisioning in nfv-enabled enterprise datacenter networks. In *Proceedings of the 12th Conference on International Conference on Network and Service Management*, CNSM 2016, page 153–159, Laxenburg, AUT, 2016. International Federation for Information Processing.

[22] Vincenzo Eramo, Emanuele Miucci, Mostafa Ammar, and Francesco Giacinto Lavacca. An approach for service function chain routing and virtual function network instance migration in network function virtualization architectures. *IEEE/ACM Transactions on Networking*, 25(4) :2008–2025, 2017.

[23] Long Qu, Chadi Assi, Maurice J. Khabbaz, and Yinghua Ye. Reliability-aware service function chaining with function decomposition and multipath routing. *IEEE Transactions on Network and Service Management*, 17(2) :835–848, 2020.

[24] Sevil Dräxler and Holger Karl. Specification, composition, and placement of network services with flexible structures. *Int. J. Netw. Manag.*, 27(2), 2017.

[25] Yongzheng Jia, Chuan Wu, Zongpeng Li, Franck Le, Alex Liu, Zongpeng Li, Yongzheng Jia, Chuan Wu, Franck Le, and Alex Liu. Online scaling of nfv service chains across geo-distributed datacenters. *IEEE/ACM Trans. Netw.*, 26(2) :699–710, apr 2018.

[26] Chen Tian, Ali Munir, Alex X. Liu, Jie Yang, and Yangming Zhao. Openfunction : An extensible data plane abstraction protocol for platform-independent software-defined middleboxes. *IEEE/ACM Trans. Netw.*, 26(3) :1488–1501, jun 2018.

[27] Xianyong Yin and Yan Ma. Aggregation service function chain mapping plan based on beetle antennae search algorithm. In *Proceedings of the 2nd International Conference on Telecommunications and Communication Engineering*, ICTCE 2018, page 225–230, New York, NY, USA, 2018. Association for Computing Machinery.

[28] Xiaojun Shang, Zhenhua Liu, and Yuanyuan Yang. Network congestion-aware online service function chain placement and load balancing. In *Proceedings of the 48th International Conference on Parallel Processing*, ICPP '19, New York, NY, USA, 2019. Association for Computing Machinery.

[29] Rui Xia, Haipeng Dai, Jiaqi Zheng, Rong Gu, Xiaoyu Wang, and Guihai Chen. Safe : Service availability via failure elimination through vnf scaling. In *Proceedings of the 48th International Conference on Parallel Processing*, ICPP '19, New York, NY, USA, 2019. Association for Computing Machinery.

[30] Yilei Lin, Ting He, Shiqiang Wang, Kevin Chan, and Stephen Pasteris. Looking glass of nfv : Inferring the structure and state of nfv network from external observations. *IEEE/ACM Trans. Netw.*, 28(4) :1477–1490, aug 2020.

[31] ETSI. Network functions virtualisation – introductory white paper. `https://portal.etsi.org/nfv/nfv_white_paper.pdf`, last accessed : 26.03.2023).

[32] ETSI - site web. liste des membres de l'etsi. `https://www.etsi.org/membership`, last accessed : 27.07.2023).

[33] ETSI - site web. liste des membres de l'etsi nfv group. `https://portal.etsi.org/TB-SiteMap/NFV/NFV-List-members`, last accessed : 27.07.2023).

[34] Tung-Wei Kuo, Bang-Heng Liou, Kate Ching-Ju Lin, and Ming-Jer Tsai. Deploying chains of virtual network functions : On the relation between link and server usage. *IEEE/ACM Trans. Netw.*, 26(4) :1562–1576, aug 2018.

[35] Arif Husen, Muhammad Hasanain Chaudary, and Farooq Ahmad. A survey on requirements of future intelligent networks : Solutions and future research directions. *ACM Comput. Surv.*, 55(4), nov 2022.

[36] R. Martínez, L. Vettori, J. Baranda, J. Mangues-Bafalluy, E. Zeydan, and B. Bakhshi. Resource abstractions in nfv management and orchestration : Experimental evaluation. *IEEE Transactions on Network and Service Management*, 20(1) :608–624, 2023.

[37] Gabriella Colajanni, Patrizia Daniele, Laura Galluccio, Christian Grasso, and Giovanni Schembra. Service chain placement optimization in 5g fanet-based network edge. *IEEE Communications Magazine*, 60(11) :60–65, 2022.

[38] Anna Engelmann and Admela Jukan. A combinatorial reliability analysis of generic service function chains in data center networks. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 6(3), dec 2021.

[39] Jing Chen, Jia Chen, and Hongke Zhang. Drl-qor : Deep reinforcement learning-based qos/qoe-aware adaptive online orchestration in nfv-enabled networks. *IEEE Transactions on Network and Service Management*, 18(2) :1758–1774, 2021.

[40] Thang Le Duc, Rafael García Leiva, Paolo Casari, and Per-Olov Östberg. Machine learning methods for reliable resource provisioning in edge-cloud computing : A survey. *ACM Comput. Surv.*, 52(5), sep 2019.

[41] Vajiheh Farhadi, Fidan Mehmeti, Ting He, Thomas F. La Porta, Hana Khamfroush, Shiqiang Wang, Kevin S. Chan, and Konstantinos Poularakis. Service placement and request scheduling for data-intensive applications in edge clouds. *IEEE/ACM Trans. Netw.*, 29(2) :779–792, apr 2021.

[42] Ashalatha Kunnappilly, Peter Backeman, and Cristina Seceleanu. From uml modeling to uppaal model checking of 5g dynamic service orchestration. In *7th Conference on the Engineering of Computer Based Systems*, ECBS 2021, New York, NY, USA, 2021. Association for Computing Machinery.

[43] Lingnan Gao and George N. Rouskas. Congestion minimization for service chain routing problems with path length considerations. *IEEE/ACM Trans. Netw.*, 28(6) :2643–2656, dec 2020.

[44] Xi Huang, Simeng Bian, Xin Gao, Weijie Wu, Ziyu Shao, Yang Yang, and John C. S. Lui. Online vnf chaining and predictive scheduling : Optimality and trade-offs. *IEEE/ACM Transactions on Networking*, 29(4) :1867–1880, 2021.

[45] Ziyue Luo and Chuan Wu. An online algorithm for vnf service chain scaling in datacenters. *IEEE/ACM Trans. Netw.*, 28(3) :1061–1073, jun 2020.

[46] Niloy Saha, Mohammad Zangooei, Morteza Golkarifard, and Raouf Boutaba. Deep reinforcement learning approaches to network slice scaling and placement : A survey. *IEEE Communications Magazine*, 61(2) :82–87, 2023.

[47] Satyam Agarwal, Francesco Malandrino, Carla Fabiana Chiasserini, and Swades De. Vnf placement and resource allocation for the support of vertical services in 5g networks. *IEEE/ACM Trans. Netw.*, 27(1) :433–446, feb 2019.

[48] Quang-Trung Luu, Sylvaine Kerboeuf, Alexandre Mouradian, and Michel Kieffer. A coverage-aware resource provisioning method for network slicing. *IEEE/ACM Trans. Netw.*, 28(6) :2393–2406, dec 2020.

[49] Yuchen Gan, Xin Su, Chang Choi, and Zhou Zhou. Clustered nfv service chaining scheme for ocean observations. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, RACS '19, page 175–180, New York, NY, USA, 2019. Association for Computing Machinery.

[50] Albert Mestres, Alberto Rodriguez-Natal, Josep Carner, Pere Barlet-Ros, Eduard Alarcón, Marc Solé, Victor Muntés-Mulero, David Meyer, Sharon Barkai, Mike J. Hibbett, Giovani Estrada, Khaldun Ma'ruf, Florin Coras, Vina Ermagan, Hugo Latapie, Chris Cassar, John Evans, Fabio Maino, Jean Walrand, and Albert Cabellos. Knowledge-defined networking. *SIGCOMM Comput. Commun. Rev.*, 47(3) :2–10, sep 2017.

[51] Jane Yen, Jianfeng Wang, Sucha Supittayapornpong, Marcos A. M. Vieira, Ramesh Govindan, and Barath Raghavan. Meeting slos in cross-platform nfv. In *Proceedings of the 16th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '20, page 509–523, New York, NY, USA, 2020. Association for Computing Machinery.

[52] Jérôme François, Alexander Clemm, Dimitri Papadimitriou, Stenio Fernandes, and Stefan Schneider. Research Challenges in Coupling Artificial Intelligence and Network Management. Internet-Draft draft-irtf-nmrg-ai-challenges-01, Internet Engineering Task Force, July 2023. Work in Progress.

[53] Amazon EC2. Instances types. `https://aws.amazon.com/fr/ec2/instance-types/`, last accessed : 16.07.2023).

[54] Amazon EC2. Tarification d'amazon ec2 à la demande. `https://aws.amazon.com/fr/ec2/pricing/on-demand/`, last accessed : 16.07.2023).

[55] Microsoft Azur. Tarification de microsoft azure à la demande. `https://azure.microsoft.com/fr-fr/pricing/calculator/`, last accessed : 30.07.2023).

[56] Google Cloud. Instance type google. `https://cloud.google.com/compute/?hl=fr#section-11`, last accessed : 30.07.2023).

[57] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A. Maltz, Parveen Patel, and Sudipta Sengupta. Vl2 : A scalable and flexible data center network. *SIGCOMM Comput. Commun. Rev.*, 39(4) :51â€“62, aug 2009.

[58] Brian Daigle, Office of Industries, US international Trade Commission. Data centers around the world : A quick look. https://www.usitc.gov/publications/332/executive_briefings/ebot_data_centers_around_the_world.pdf, last accessed : 31.07.2023).

[59] Le Soir. Google en 15 chiffres fous. https://references.lesoir.be/article/google-en-15-chiffres-fous/, last accessed : 16.07.2023).

[60] Paul Anirban. Network function virtualization. https://learn.microsoft.com/en-us/windows-server/networking/sdn/technologies/network-function-virtualization/network-function-virtualization, last accessed : 03.08.2023).

[61] Raj Vardhman. 15 crucial data center statistics to know in 2023. https://techjury.net/blog/data-center-statistics/#:~:text=8.,the%20world%20reaches%20almost%208%2C000.&text=The%20world%20has%20approximately%208%2C000,locations%20all%20around%20the%20world., last accessed : 03.08.2023).

Chapitre 9

Acronymes

AI Artificial Intelligence

BSS Business Support Systems

DRL Deep Reinforcement Learning

EM Element Manager

ETSI European Telecommunications Standards Institute

GDPR General Data Protection Regulation

IDS Intrusion Detection System

IETF Internet Engineering Task Force

ILP Integer Linear Programming

IoT Internet of Things

IPS Intrusion Prevention System

IRTF Internet Research Task Force

LB Load Balancer

MANO Management ANd Orchestration

MILP Mixed-Integer Linear Programming

ML Machine Learning

NAT Network Address Translation

NF Network Function

NFV Network Functions Virtualisation

NFVI Network Functions Virtualisation Infrastructure

NFV-O NFV Orchestrator

OSM Open Source MANO

OSS Operational Support Systems

RFC Request For Comments

SDN Software-Defined Networking

SF Service Function

SFC Service Function Chain

SLA Service Level Agreement

SLO Service Level Objective

VIM Virtualized Infrastructure Manager

VNF Virtual Network Function(s)