ÉCOLE DOCTORALE GALILÉE
Université Sorbonne Paris Nord

Contrat doctoral – ED Galilée

Titre du sujet : Continuum Cloud-Edge : vers une orchestration dynamique et de bout en bout des micro-services

- Unité de recherche : LIPN- L2TI
- Discipline : Informatique
- Direction de thèse : Christophe Cérin, Gladys Diaz, Khaled Boussetta
- Contact : Christophe Cerin <christophe.cerin@univ-paris13.fr>
- Domaine de recherche : Cloud, Edge and Fog computing
- Mots clés : Continuum Cloud-Edge, réseaux, micro-services, orchestration, Service Level Agreement (SLA)

Contexte scientifique et problématique

L'émergence des architectures *Cloud Continuum*, intégrant les capacités de traitement distribué du Cloud avec les avantages de la connectivité réseau (par exemple 5G/6G), ouvre de nouvelles opportunités pour le déploiement d'applications évolutives et performantes. Cependant, l'orchestration efficace des microservices au sein de ces environnements présente des défis, notamment en termes de gestion des ressources, de latence réseau et de sécurité. Ainsi est apparue la notion de Fog computing que l'on peut voir comme une entité entre le Cloud et l'Edge.

Cette thèse s'inscrit dans le cadre et le prolongement du projet FogSLA-Antillas, où l'orchestration des architectures Cloud Continuum est l'un des défis scientifiques. Ce projet qui implique plusieurs partenaires industriels et académiques se propose de développer une plateforme de services Fog. Cette plateforme reposera à terme sur deux innovations : la description des exigences unitaires dans un contrat *dynamique* et la prise en compte de contraintes multicritères pour les interfaces d'orchestration. L'optimisation de l'orchestration de bout en bout des services Fog est l'une des principales contributions attendues.

Défis scientifiques

Dans l'environnement de la plateforme FogSLA, plusieurs microservices coexistent et ce sur des environnements hétérogènes. Une orchestration de service de bout en bout est nécessaire pour maintenir un contrôle continu des services déployés selon la Qualité de Service (QoS) demandée par les cas d'utilisation. Ce point nous amène à la question de savoir comment exprimer dynamiquement les demandes de QoS à partir de cas d'utilisation dans un SLA, qui pourrait ensuite être utilisé pour définir des objectifs d'orchestration de bout en bout (SLO).

Concernant l'orchestration, lorsqu'un objectif est fixé, il existe plusieurs façons de l'atteindre, en passant par le déploiement de plusieurs services, et en général chaque étape introduit des propriétés qui peuvent être plus ou moins attractives pour l'objectif. Une difficulté est que ces propriétés sont découvertes en cours de route et doivent être appliquées dynamiquement. Examiner les problèmes de SLA pour la 5G du point de vue des SLO et de la dynamicité dans l'orchestration ainsi que le cycle de vie des services est encore une problématique ouverte.

La décision sur la manière de connecter les différents niveaux (Fog, Edge, Cloud) dépend d'un scénario technologique spécifique. Ainsi, différentes topologies et technologies réseau peuvent coexister pour connecter les « nœuds » Fog aux mêmes ou à différents niveaux. Comment maintenir cette connectivité et comment adapter dynamiquement le routage en fonction des objectifs SLA/SLO est également une question à traiter.

Objectifs de la thèse

La recherche se concentrera sur le développement de mécanismes d'orchestration de bout en bout efficaces capables de déployer, gérer et mettre à l'échelle dynamiquement des micro-services sur des nœuds du cloud et en périphérie (Edge) en fonction des demandes de charge de travail, de la disponibilité des ressources et des exigences de qualité de service. Pour relever ce défi, le concept de "SLA dynamique" sera étudié pour introduire et orchestrer des configurations flexibles. Les travaux seront développés autour deux aspects majeurs :

(1)Optimisation de la planification de conteneurs : les techniques de planification de conteneurs sur le marché limitent les possibilités d'optimisation à une approche standardisée basée sur un seul critère. L'objectif ici est de poursuivre la mise en œuvre de solutions d'optimisation de la planification de conteneurs dans un contexte multi-critère. La solution clé proposée sera adaptée à Kubernetes, dans un environnement de Fog computing, pour réduire le coût du placement des conteneurs tout en garantissant une optimisation des performances en termes de temps de calcul. Avec la notion de calcul écologique en tête, nous proposons d'étudier le problème de définir un modèle d'optimisation pour les requêtes de données sur les réseaux de véhicules dans les environnements Fog/Edge. Ce modèle, qui ne devra contenir que des notions « dynamiques » (en ligne) parce que les données arrivent continuellement devra ajuster automatiquement le nombre de nœuds Fog actifs en temps réel.

(2)Connectivité réseau et routage : ce point concerne le maintien de la connectivité entre les niveaux Edge et Fog, et les niveaux Fog et Cloud. La connectivité entre les différents niveaux de la hiérarchie Fog/Cloud peut être possible en utilisant plusieurs technologies réseau, y compris des réseaux câblés et sans fil. Dans une architecture basée sur le Edge-Cloud Continuum, le nœud Fog devient une unité fonctionnelle incorporant des capacités de calcul, de stockage et offrant des fonctionnalités réseau. Nous proposons d'étudier de nouvelles procédures pour la fourniture de services numériques Fog et 5G, basées sur l'approche de virtualisation (NFV - Network Function Virtualization). Le niveau Fog permettant de définir des mécanismes de routage et des aspects de contrôle et le niveau de Cloud permettant de définir l'orchestration End-to-End (E2E). Nous proposons d'étudier l'introduction de l'approche SDN (Software Defined Networking) au niveau de Fog. Ici, nous devons considérer les problèmes de SLA pour la 5G du point de vue des SLOs et le déploiement dynamique des services réseau au niveau du Fog. Dans ce contexte, le routage basé sur l'intention en intégrant, par exemple, le contrôleur ONOS, pourrait être considéré. Quelques pistes de recherche connexes sont les suivantes : Routage centré sur les données énergétiquement efficace, routage multi-objectif orienté application et routage de centre de données basé sur l'intention.

Plan de travail approche méthodologique

1. La première étape consiste à réaliser l'état de l'art dans les domaines associés à cette étude.
2. Il sera ensuite nécessaire de définir les cas d'utilisation à traiter, les micro-services à considérer, les critères d'optimisation pour le déploiement des micro-services et des conteneurs. Il en va de même pour les objectifs de gestion attendus au niveau de l'architecture et pour les aspects réseau et de routage. Ainsi, la plateforme permettra, selon ces cas d'utilisation, de garantir des niveaux d'exécution, sur des environnements hétérogènes, ces calculs reposant sur une observation continue et un contrôle entre la QoE et la définition de la QoS.
3. Nous procéderons également à l'implémentation des solutions choisies afin d'évaluer leurs performances.

Bibliographie

[1] D. Zeng, N. Ansari, M. -J. Montpetit, E. M. Schooler and D. Tarchi, "Guest Editorial: In-Network Computing: Emerging Trends for the Edge-Cloud Continuum," in IEEE Network, vol. 35, no. 5, pp. 12-13, September/October 2021, doi: 10.1109/MNET.2021.9606835.
[2] E. Kapassa, M. Touloupou, A. Mavrogiorgou, D. Kyriazis, "5G & SLAs: Automated proposition and management of agreements towards QoS enforcement", 21st IEEE Conference on Innovation in Clouds, Internet and Networks (ICIN), Paris, France, 2018.
[3] I. Mesogiti et al., "Network Services SLAs over 5G Infrastructure Converging Disaggregated Network and Compute Resources," 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Barcelona, 2018, pp. 1-5. doi: 10.1109/CAMAD.2018.8514989
[4] M. Touloupou, E. Kapassa, C. Symvoulidis, P. Stavrianos and D. Kyriazis, "An Integrated SLA Management Framework in a 5G Environment," 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), Paris, France, 2019, pp. 233-235. doi: 10.1109/ICIN.2019.8685916
[5] K. Fu, W. Zhang, Q. Chen, D. Zeng and M. Guo, "Adaptive Resource Efficient Microservice Deployment in Cloud-Edge Continuum," in IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 8, pp. 1825-1840, 1 Aug. 2022, doi: 10.1109/TPDS.2021.3128037.