Cloud Edge Continuum to support emerging network services that require low latency and high bandwidth usage

1. Context

Cloud Edge Continuum (CEC) [1] is the natural evolution of the cloud computing paradigm, as new generations of service such as VR/AR, cloud gaming, autonomous driving are envisioned. This is motivated by the widespread use of mobile devices, and the deployment of new mobile network generation (5G) that will improve communications by providing higher bandwidth and low-latency. CEC aims to unify computing platforms that run at the edge close to end-users to reduce low-latency required by services like AR/VR, with centralized cloud that runs more persistent and delay-tolerant services. It should be noted that Edge resources include MEC as well as end-device computing. CEC has gained momentum recently with the common interests shown by network operators and cloud operators, example is the agreements made between AWS and telco operators such as Verizon in US [2]. On one hand, network operators want to open their network to host edge platforms and allow application developer to take advantage of this environment to deploy low latency application and services; on the other hand cloud operators want to take advantage of their well-known experience to enable application developers to run everywhere their application including low-latency and bandwidth eager services. Meanwhile, serverless architecture [3] and Function as a Service (FaaS) [4] are new concepts that allow to divide a service or an application into several micro-services that can run independently, and can be flexibly deployed and run on different platforms. Combined with CEC, FaaS and serverless concepts will enable the notion of service or application continuum, where part of the service can be deployed at the end-user device, part at the edge and part at the centralized cloud. But location where part of a service has to be deployed is dependent on the application's required latency, bandwidth requirements, and the resource optimization usage. Here we can mention the case of cloud gaming [5], where part is executed at the user device, and some functions are offloaded at the edge or centralized cloud.

CEC is in infancy, and many challenges remain and need to be addressed:

- How to build a continuum service that can transparently run in different locations?
- How to manage and orchestrate in a unified way cloud and edge resources? Knowing that different actors may be involved; network providers, cloud providers, and end-user devices.

In this thesis project, we aim to clarify and answer those questions, by studying and proposing architectural and algorithmic solution to enable Cloud Edge continuum and answer the above questions and challenges.

2. Objectives

The thesis project aims to address the challenge related to the deployment of CEC aiming at enabling novel services, particularly those requiring low latency communication and computation, and high bandwidth usage. To reach these objectives, during this thesis we will address the following challenges and issues:

Cloud and Edge operator collaboration

Cloud operators manage mainly big data centers, using their own tools and mechanisms to propose their services to tenants and end-users, while edge cloud is managed by network operators, as the edge resources are deployed by the network operators in its premises to be close to the end users. Therefore, there is a need to align cloud operators and network operators management and orchestration tools to be able to deploy CEC. This will allow the cloud operators to continue proposing their services while covering edge resources, and enabling the deployment of services in the network operator premises; the network operators to extend their business to the deployment of third tiers applications and services on their infrastructure to maximize their profit, and take benefit from the well know experience of cloud operators. However, this comes with several challenges to resolve, such as the definition of new protocols and interfaces to integrate network operator edge platform, which are mainly based on the MEC standard, with the centralized cloud management and orchestration system.

Edge resources discovery in a heterogenous environment

In the CEC a service should be divided into micro-services introducing the need of edge resources discovery mechanisms, particularly in the envisioned heterogeneous environment. Indeed, some micro-services can serve more than one micro-service, and hence may be shared to efficiency optimize the computing resources. Example of such a shared micro-service could be a Machine Learning (ML) function, which is called by different functions to realize a prediction or provide answer to an action. Therefore, it is important that each edge has to announce the available shared resources and functions that can be used by other functions or micro-services. In this context, algorithms based on web services can be explored to dynamically discover edge resources and services availability.

Micro-service placement in CEC

When a service has to be deployed on the CEC, it has to be divided into micro-services which will be deployed according to the service requirement in terms of latency, bandwidth consumption and usage, and the computing usage. Therefore, optimization techniques need to be investigated to find the optimal placement of micro-services in the CEC. Specifically, the derived algorithms need to take into consideration the dynamic of the micro-services, as in the context of FaaS, a function is executed only when it is called; meaning that its execution depend principally on the requests that come from the other micro-services. Predictive solutions based on ML can be also explored to address this issue.

User mobility and their impact on the micro-service placement

Once the service is deployed, the mobility of the end user may impact the performances of the run service, since users may move to a location where the edge connectivity is not optimal, i.e., another edge is available and can ensure better latency. Accordingly, there is a need to devise algorithms that continually track the quality of the connectivity of users and network resources, and ensure that service is running as expected. Solutions based on Reinforcement Learning can be explored to migrate micro-services among edges to ensure the best connectivity for the service and fulfil its requirements in terms of latency and bandwidth consumption.

3. Organization

The PhD project is divided into three phases:

- The first phase, from M0 to M6, will be dedicated to related work on cloud and edge computing in terms of architecture, management and networking, serverless computing model as well as Function as a Service (FaaS). Also, state of art on CEC needs to be established. This phase will allow a better understanding of the challenges related to CEC.

- The second phase, from M6 to M30, will consist of devising algorithms and mechanisms to address the CEC challenges. Architecture and system oriented solutions are expected, in addition to more algorithmic and formal solutions. The proposed solutions need first to be evaluated via computer simulation, but some contributions will be selected and evaluated via a proof of concept (PoC). The latter will rely on open-source tools such as OpenAirInterface (OAI) and the MEC platform of EURECOM.
- The last phase, from M30 to M36, will be dedicated to the PhD document and the preparation of the final defense.

For the dissemination activities, we aim to publish and demonstrate the devised works in peer-reviewed conferences, such as IEEE ICC, Globecom, and ACM Cloud Computing. During the final year, one or two publications will be submitted to peer-reviewed journals

Contact:

Prof. Adlen Ksentini (adlen.ksentini@eurecom.fr)

References:

[1] L. Baresi, D. Mendonça, M. Garriga, S. Guinea, “A Unified Model for the Mobile-Edge-Cloud Continuum”, in ACM Transactions on Internet Technology 2019

[2] https://aws.amazon.com/fr/wavelength/

[3] L. Baresi, D. Mendonça, “Towards a Serverless Platform for Edge Computing”, IEEE conference on Fog Computing (ICFC) 2019

[4] T. Pfandzelter, D. Bermbach, “tinyFaaS: A Lightweight FaaS Platform for Edge Environments”, IEEE conference on Fog Computing (ICFC) 2020

[5] F. Messaoudi, A. Ksentini, G. Simon, P. Bertin, “Performance Analysis of Game Engines on Mobile and Fixed Devices”, ACM Transactions on Multimedia Computing Communications and Applications 13(4):1-28