Provisioning of Applications with Network Requirements on Cloud-Edge Continuum

Paulo Roberto Albuquerque[1], **Guilherme Piêgas Koslovski**[1]

[1]Graduate Program in Applied Computing
Santa Catarina State University – Joinville – SC – Brazil

Abstract. *New challenges regarding the users mobility and network latency are observed in the provisioning of distributed applications atop Cloud Continuum scenarios. This requires a analysis from the perspectives of users and service providers. This article proposes a way to create provisioning policies based on user mobility and mass demands, while also analyzing the individual Services that compose some Application, analyzing their needs and restrictions.*

1. Introduction

Even though Cloud Computing dates back to last century, it truly started to take shape in the 2000s, when big companies started to heavily invest in building Data Center infrastructure to support the hosting of applications, and renting this infrastructure to others in a plenitude of paradigms. Currently, many cloud-hosted applications are often considered a single and monolithic functionality or task executor. However, that is rarely the case, most often, applications are composed by big architectures of services and components, each with a specific job/function. This method of constructing applications is very beneficial to the Cloud Native way of deploying something, that is because it allows each components to scale individually according to its workload. This of course comes at the expense of everything being deployed in a Data Center somewhere. Now, typical Data Center deployments often don't suffer from very high latencies and other problems that may arise from a end-user being very distant in relation to the computing source. However, in certain scenarios, the real-timeness of some components is of utmost importance, such as critical operational sensors accessing or communicating with an Application through weak networks, or users accessing through mobile devices [Mahmoudi et al. 2018]. In these scenarios, the lower latencies and bandwidth economicality of the Edge Computing paradigm can be very beneficial. Unfortunately, Edge Computing has some very serious and known limitations in regards to processing power and storage amounts available [Luo et al. 2021]. It is limited to what the edge hardware has to offer, which considering its power efficiency and cost effectiveness restrictions, is very frequently not enough for certain parts of applications, so how could any application ever have the optimal set of resources if it can either be on the Cloud, or on the Edge [Garcia Lopez et al. 2015]. That is the exact reasoning behind the Cloud-Edge Continuum paradigm. When there is an application in a very specific scenario which could benefit from both the characteristics of Cloud Computing and Edge Computing, there is a very large interest in trying to provision it in both of theses networking substrates. The classification of applications in regards to its composing services and components seems to be a not so explored topic in this research realm, and also would benefit the formulation of a provisioning policy in order to determine how to optimize certain metrics related to network latency and computing resources usage/efficiency, as this could positively improve the perceived performance of the application itself for end-users, depending on the analyzed metrics.

2. Proposal

Trying to leverage the full capabilities of the whole Cloud-Edge Continuum by consciously allocating applications throughout the whole substrate is a promising approach to Services Provisioning. There are two main aspects that need attention in this scenario, the first one is correctly managing the Edge resources and allocating services in a as-close to optimal way as possible. This needs to be done via a provisioning policy, which can vary wildly in objective. One of the ways to try to predict the correct distribution of resources is to couple the users and their needs to the corresponding services they consume, and where they are allocated as opposed to where they could be. The other aspect is determining which of the services benefit from being on the Edge, opposed to the ones that have some specific need or restriction, and require being on the Cloud [Fu et al. 2021]. There are a couple of ways of making this service substrate distinction, one is to make historical data of many application the basis of analysis, and the other is making each application identify their restriction and needs respectively. In this ongoing research, we proposes to utilize user mobility data through temporal and geographical axes to predict possible reallocations of services in Edge, while also leveraging concepts such as Infrastructure as Code to determine which services have special necessities and need to be allocated in the Cloud, with a focus on data streaming/processing Applications and Workloads. One of the main challenges is how to run experiments to a proper scale, taking into consideration the lack of access to a real substrate containing both enough Edge resources to demonstrate that proper provisioning can lead to gain in performance and other metrics, and also access to a Cloud infrastructure to properly leverage high performance computing. For this, the usage of a Edge Computing Simulator [Souza et al. 2023] will be done, accompanying with the usage of real datasets of geographical distribution of Edge servers, base stations, antennas, and also, user mobility datasets and simulated models. We propose to pursue a classification of Applications such that the needs and restrictions for each of their services becomes clear and can be expressed in a declarative way, and also of a Edge substrate provisioning policy, to better accommodate with ever changing user needs and demands.

References

Fu, K., Zhang, W., Chen, Q., Zeng, D., and Guo, M. (2021). Adaptive resource efficient microservice deployment in cloud-edge continuum. *IEEE Transactions on Parallel and Distributed Systems*, 33(8):1825–1840.

Garcia Lopez, P., Montresor, A., Epema, D., Datta, A., Higashino, T., Iamnitchi, A., Barcellos, M., Felber, P., and Riviere, E. (2015). Edge-centric computing: Vision and challenges.

Luo, Q., Hu, S., Li, C., Li, G., and Shi, W. (2021). Resource scheduling in edge computing: A survey. *IEEE Communications Surveys & Tutorials*, 23(4):2131–2165.

Mahmoudi, C., Mourlin, F., and Battou, A. (2018). Formal definition of edge computing: An emphasis on mobile cloud and iot composition. In *2018 Third international conference on fog and mobile edge computing (FMEC)*, pages 34–42. IEEE.

Souza, P. S., Ferreto, T., and Calheiros, R. N. (2023). Edgesimpy: Python-based modeling and simulation of edge computing resource management policies. *Future Generation Computer Systems*, 148:446–459.