Multi-Time-Scale Distribution of Latency-Sensitive Tasks in Mobile Edge Computing Networks

Guilherme Iecker Ricardo[1], Amal Benhamiche[1], Nancy Perrot[1], Yannick Carlinet[1]

Orange Innovation, Châtillon, France

{guilhermeiecker.ricardo, amal.benhamiche, nancy.perrot, yannick.carlinet}@orange.com

Keywords : *Edge Computing, Resource Allocation, Quality of Service, Operations Research.*

1 Introduction

The world has witnessed an intense digital transformation with the emergence of disruptive systems, such as pervasive augmented reality and enhanced virtual social networking. These systems share two common characteristics: (i) low latency tolerance and (ii) high data communication throughput demand. For example, for virtual and augmented reality systems, guaranteeing sufficiently low latency provides the right level of responsiveness at the same time that higher data throughput provides a better immersion and more fluid experience. It is an important challenge in 6G design to provide a network infrastructure that is able to support systems with such strict requirements.

Edge Computing has been considered as a key enabling technology for 6G networks to meet a series of use cases' requirements, particularly latency mitigation. Edge-Computing-enabled mobile networks are often referred to as Mobile Edge Computing (MEC) networks. For the aforementioned systems deployed on top of MEC networks, Quality of Service (QoS) may be defined as a function of the achieved data throughput, while a parallel effort is dedicated to guarantee an end-to-end communication latency with respect to a certain tolerance. In this work, we investigate how to maximize the overall QoS in a time-evolving network with limited available computational and communication resources.

2 Problem Description

Consider a network consisting of Mobile Users (MUs) uniquely and statically associated with Access Points (APs), which, in turn, are connected to the Internet through the Edge network. The Edge network is composed of interconnected Mobile Edge Hosts (MEH), i.e., APs, intermediate switches/routers, gateways, etc., that are equipped with computational capacity to process data and are able to host and provide requested tasks.

There is a centralized intelligence aware of the entire network infrastructure, which we henceforth refer to as Mobile Edge Orchestrator (MEO). The MEO is responsible for deciding how MUs' requested tasks will be provided considering the underlying network's limited capacities. The MEO's decision takes three major components into consideration: (i) task deployment and request provision by Mobile Edge Hosts (MEHs), (ii) definition and assignment of data flows (or routes) connecting MEHs and MUs, and (iii) requests' priority assignment, which quantifies the achieved performance level as well as its corresponding amount of consumed resources. We adopt a generalization of the mathematical model proposed in [1] in its so-called flow formulation.

We observe the system for T time slots of duration Δ that we call *management* intervals. Each management interval is, in turn, split into T' sub-slots of duration δ that we call *service*

intervals. Given that the MEO is not aware of which tasks will be requested, it must find a provision setup $\mathbf{x}$ that maximizes the expected QoS at every service interval, i.e.,

$$\begin{aligned}\frac{1}{TT'}\sum_{i=1}^{TT'}\mathbb{E}\left[\mathrm{QoS}_{\mathcal{R},i}(\mathbf{x})\right] &= \frac{1}{TT'}\sum_{i=1}^{TT'}\sum_{R\subseteq\mathcal{U}\times\mathcal{T}}\mathrm{QoS}_{R,i}(\mathbf{x})\cdot\mathbb{P}(\mathcal{R}=R)\\ &= \frac{1}{TT'}\sum_{i=1}^{TT'}\sum_{R\subseteq\mathcal{U}\times\mathcal{T}}\mathrm{QoS}_{R,i}(\mathbf{x})\left[\prod_{r\in\mathcal{U}\times\mathcal{T}}\mathbb{1}_{r\in R}(2\lambda_{i-1}^{r}-1)+1-\lambda_{i-1}^{r}\right],\end{aligned} \tag{1}$$

where λ_i^r, for any $r=(u,t)\in\mathcal{U}\times\mathcal{T}$, represents the frequency at which MU u requested task t up to service interval i. Note that the QoS is a function of the random variable $\mathcal{R}$, which captures the uncertainty of the actual set of requests. The resulting provision setup must also ensure that the *expected resource utilization* satisfy a given threshold. We remark that, for the same provision setup, the *instantaneous resource utilization* at every service interval may result in resource overuse and/or constraint violation, leading to sub-optimal or even infeasible solutions in practice.

3 Methodology

In order to work around the negative effects of instantaneous resource utilization, we propose to decompose the original optimization model into two sub-problems that will be handled in different time scales. We refer to this approach as the Multi-Time-Scale Task Distribution Problem (Multi-TS TDP). Our approach was inspired by the work introduced in [2]

At the beginning of each management interval, the MEO must decide (i) at which MEHs each task will be implemented and (ii) what is the set of available network flows. These decisions are based on the history of solutions and requests from previous management intervals. We call it the *Management* sub-problem and, because it takes place in a longer interval, it may have enough time to be solved to optimality using classic integer programming methods.

Given a fixed provision setup, at every subsequent service interval, the MEO will handle the set of posed requests on-the-fly. We call it the *Service* sub-problem and it will be focused on choosing (i) provider, (ii) flow, and (iii) priority for each request. The Service sub-problem must be solved within the short time interval available before new requests are posed. Therefore, faster approaches are more suitable to address the problem, for example the approximate heuristics proposed in [3].

Acknowledgment

The authors are with the Department of Mathematical Models for Optimization and Performance Evaluation, Orange Innovation, Châtillon, 92012 France. This work is supported by the European Union H2020 Project DEDICAT 6G under grant number 101016499.

References

[1] G. Iecker Ricardo, A. Benhamiche, N. Perrot, and Y. Carlinet, "Latency-constrained task distribution in multi-access edge computing systems," in *IEEE CLOUDNET 2022 - IEEE International Conference on Cloud Networking*, 2022.

[2] R. Zhou, X. Wu, H. Tan, and R. Zhang, "Two time-scale joint service caching and task offloading for uav-assisted mobile edge computing," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 1189–1198.

[3] G. Iecker Ricardo, A. Benhamiche, N. Perrot, and Y. Carlinet, "Heuristic distribution of latency-sensitive tasks in multi-access edge computing systems," in *IEEE GLOBECOM 2022 - IEEE Global Communications Conference*, 2022.