Farris, Ivan; Taleb, Tarik; Bagaa, Miloud; Flick, Hannu

Optimizing service replication for mobile delay-sensitive applications in 5G edge network

Published in:
2017 IEEE International Conference on Communications, ICC 2017

DOI:
10.1109/ICC.2017.7997282

Published: 28/07/2017

Document Version
Peer reviewed version

Please cite the original version:
Farris, I., Taleb, T., Bagaa, M., & Flick, H. (2017). Optimizing service replication for mobile delay-sensitive applications in 5G edge network. In 2017 IEEE International Conference on Communications, ICC 2017 [7997282] (IEEE International Conference on Communications). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICC.2017.7997282

Optimizing Service Replication for Mobile Delay-sensitive Applications in 5G Edge Network

Ivan Farris[1], Tarik Taleb[1], Miloud Bagaa[1], and Hannu Flick[2]

[1]School of Electrical Engineering, Aalto University, {ivan.farris, tarik.taleb, miloud.bagaa}@aalto.fi
[2] Nokia Bell Labs, Finland, hannu.flinck@nokia-bell-labs.com

***Abstract*—Extending cloud infrastructure to the Network Edge represents a breakthrough to support delay-sensitive applications in next 5G cellular systems. In this context, to enable ultra-short response times, fast relocation of service instances between edge nodes is required to cope with user mobility. To face this issue, proactive service replication is considered a promising strategy to reduce the overall migration time and to guarantee the desired Quality of Experience (QoE). On the other hand, the provisioning of replicas over multiple edge nodes increases the resource consumption of constrained edge nodes and the relevant deployment cost. Given the two conflicting objectives, in this paper we investigate different optimization models for proactive service migration at the Network Edge, which can exploit prediction of user mobility patterns. In particular, we define two Integer Linear Problem optimization schemes, which aim at respectively minimizing the QoE degradation due to service migration, and the cost of replicas' deployment. Performance evaluation shows the effectiveness of our proposed solutions.**

I. Introduction

Over the last years both network and cloud research communities have greatly focused on the development of efficient frameworks to extend the cloud from remote data-centers to the network edge [1]. This approach could increase the support of real-time cloud applications with strict requirements in terms of latency, such as mobile gaming, augmented reality, and Tactile Internet [2]. To promote the wide adoption of edge infrastructure, several industrial organizations are working to define standards and enable interoperability, such as ETSI Mobile Edge Computing (MEC) [3] and Open Fog Consortium [4]. However, manifold challenges still remain open.

In a distributed edge environment, a key issue is represented by user mobility, which may cause significant Quality of Experience (QoE) degradation, even in small scenarios, as reported in [5]. Indeed, when a user moves under the coverage of a different edge node, he is forced to access the service instance running in the previous serving edge node for the total migration time, i.e., the time required to migrate the service instance in the attached edge node. Since the links between edge nodes usually present high latency, forwarding the request to another edge node can notably increase service response time, and so drastically reduce the user QoE.

To enable fast migration of application instances between edge nodes, service replication mechanism can be adopted [6]. Indeed, the proactive deployments of relevant application replicas in near edge nodes can make the migration seamless to the end user, who will connect to a ready replica instance after moving to a different serving edge node. This approach can be especially promising for novel applications based on microservice architecture, which relies on moving application states into back-end data systems [7] [8] [9]. As investigated in [10] [11], proactive service migration solutions for container-based applications present several benefits compared to classic reactive migration systems. By enabling appropriate synchronization between back-end storage and thus providing fast access to on-demand state, extremely low-latency interaction between the processing service and relevant data store allows for guaranteeing adequate application performance. However, service replication involves additional cost of deployment in terms of processing and storage resources for the replicas allocation, and bandwidth for guaranteeing periodical synchronization between relevant replicas. Accounting also that edge nodes are typically resource constrained, appropriate optimization schemes should be adopted to enable an efficient use of the service replication mechanism, especially for 5G cloud-based cellular systems.

In this paper, we formulate the problem of proactive migration approach in cellular 5G networks. Furthermore, by leveraging on prediction schemes of user mobility patterns, we propose analytic models based on Integer Linear Problem optimization to enhance the proactive migration of applications in a multi-edge environment. In particular, we define two solutions aiming at minimizing, on the one hand, the QoE degradation due to service migration, and on the other hand, the cost of proactive replication. Finally, we evaluate the proposed schemes accounting for different configurations in MEC scenarios.

This paper is structured as follows. In the next section, we browse relevant related work. In Section III, we discuss the reference system model for service migration, whereas the problem formulation is defined in Section IV. The proposed optimization solutions are described in Section V. Performance evaluation is presented in Section VI, while conclusive remarks are given in the last section.

II. Related work

Nowadays, migration of service instances represents an essential functionality in cloud distributed infrastructures [12]. Different techniques for live migration of VMs in data-centers have been developed, such as Pre-Copy and Post-Copy memory migration [13]. Several research works have been also

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

conducted to extend live VM migration over distributed cloud data-centers. In [5], a specific VM handoff method has been proposed for MEC environment to tolerate the high variability of cloudlets. Despite the remarkable efforts, the results still present significant values of total migration times, which could cause a notable degradation of QoE in MEC environment, since degraded end-to-end latency persists until the end of the operation.

The service replication approaches have been highly studied in the literature for VMs to support high availability (HA) [14]. Furthermore, CloudSpider [6] proposes combining VM replication with VM scheduling to reduce migration times due to the transfer of large VM image over low bandwidth WAN (Wide Area Network) links. Our approach is also focused on novel microservices applications and container-based virtualization [15] [16], which cope better with the limitation of resource-constrained edge nodes.

Another relevant area of research deals with Application Content Delivery Network at the edge [17], [18], [19], which investigates the deployment of web services in different edge servers, by distributing both service code and dynamic content. Whereas our study addresses real-time replication of services based on single user migration, these research work focus on the data management, since concurrent accesses could corrupt data consistency between the web servers.

Furthermore, in cellular systems, channel reservation schemes and models have been highly investigated to reduce the probability of dropped calls during user handovers [20]. To this aim, a number of channels is reserved to ensure the priority of successful handover calls. However, these available channels can be exploited interchangeably by the users which move to the same cell. This differs from our MEC scenario whereby each deployed application instance is strictly associated to a single user, accounting for the relevant data and state. Therefore, the replicas deployed in the near edge nodes are reserved only for the user which has requested the related service.

III. Service migration in MEC Environment

To meet the strict latency requirements of real-time applications [2], such as Tactile Internet, mobile gaming, and augmented reality, moving the deployment of user applications at the network edge will play a key role in next-generation cellular networks. However, the mobility of users between different edge nodes will introduce new challenges for cloud-based Telco infrastructure [21]. In this section, we consider different service migration solutions for MEC-based applications to meet the desired QoE.

A. Reactive approach

A classic reactive approach requires the migration of a service instance between different edge nodes following the mobility of its respective user. In particular, when a user moves out from the coverage area of a serving edge node (hereinafter referred to as source edge), the MEC framework needs to first locate the most suitable target edge (i.e., both in terms of geographical proximity and resource availability) to accommodate the migrating service. Once the target edge is selected, the procedure of migration for user application is performed, according to the type of application and the relevant underlying virtualization technology [22].

It is worth noticing that, in MEC environment, service continuity is not only related to the downtime (i.e., the time when the application instance is interrupted to perform the final phase of service migration), but also to the overall total migration time, which drastically impacts the user's desired QoE. Indeed, until the completion of the migration process, the user needs to interact with the serving instance in the source edge and the relevant round-trip time may exceed the tolerable delay.

B. Proactive approach

To support ultra-short latency applications, we consider proactive migration in a MEC cellular environment. The basic idea is to guarantee fast relocation by deploying multiple replicas of the user service in neighboring edge nodes, so to make services readily available if handover to a different cloudlet is necessary. In this way, the long migration time required by on-demand service relocation could be drastically reduced.

However, the proposed approach introduces additional cost per each user service. First, the replication mechanism must reserve enough resources, in terms of computation, memory, and networking to guarantee the execution of the service replicas according to the required Service Level Agreement (SLA). Moreover, the system is required to provide a synchronization system, which manages the periodical propagation of state updates between replicas. This aspect involves further cost in terms of bandwidth to deploy replicated service instances in order to guarantee periodical data synchronization.

An optimized planning of service replicas, by carefully selecting the number and location in a cellular environment, could drastically reduce the overall cost. Indeed, when considering a traditional cellular cluster topology, a naive approach is represented by distributing replicas in all edges nearby the serving cloudlet. However, this approach involves two drawbacks. On the one hand, there is a risk of wasting reserved resources in edge nodes which the user will never connect to; on the other hand, there is the risk to overload cloudlets with replicas when a large number of user requires delay-sensitive services (thus reducing the number of new effective services deployed due to lack of available resources). As a possible solution to reduce the impact of service replication, precise estimation of user mobility pattern can drastically reduce the cost of replica deployment. For instance, by exploiting information about user position and orientation, the system could perform simple predictions of future user path and replicate services only at the target edges towards the user mobility direction. In this way, the number of involved edge nodes could be almost halved. More advanced prediction of user mobility patterns can be developed by exploiting a mobile service usage cartography based on filtered historical movements [23] and leveraging

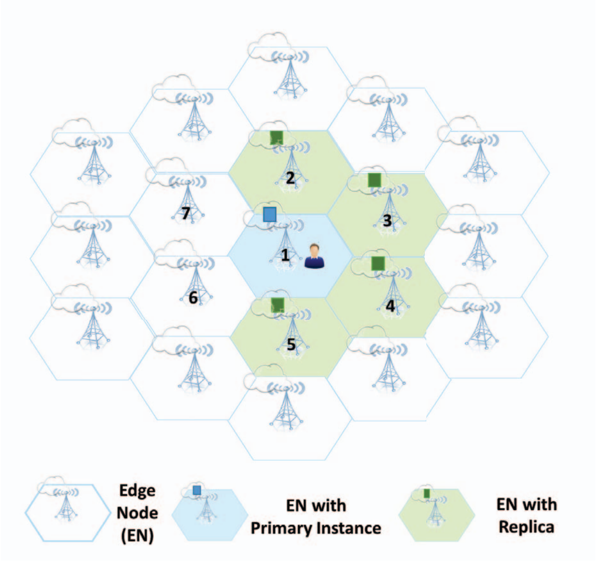


Fig. 1. Example deployment of proactive migration based on replication.

big data techniques [24]. For instance, in vehicular urban environment, highly accurate estimation of handover events along the path to destination can be modeled by accounting for the time taken to transit road segments along the path, navigation zone characteristics, and behaviour of users on specific road segments. Therefore, by integrating model able to predict the user path from source to destination [25], the MEC system could evaluate if enough resources can be reserved in the cloudlets along the user route, in order to enable extremely fast handover between the selected edge nodes and guarantee the desired QoE.

In Fig. 1, an exemplary scenario of replicas deployment for a typical cellular cluster is depicted. We assume that the user is under the coverage of edge node EN_1 and a relevant user application is running in EN_1. According to the probability of handover in nearby cells, service replicas of the same user's application are deployed in edge nodes EN_2, EN_3, EN_4, and EN_5. If the user moves to one of these edge nodes, fast service migration is guaranteed by the proactive approach, whereas if user moves to either EN_6 and EN_7 a reactive migration is required.

IV. Problem formulation

Let us assume that the MEC system is composed of a set of N Edge Nodes (ENs), named $\mathcal{N}$, where each EN is referred with an index $1 < n < N$. Let define the population of U users, named $\mathcal{U}$, that are in the MEC environment. Let us assume that each user requires only one service. Each primary instance of user application is executed in the edge node which the user is connected to. Indeed, we underline that the proposed model is designed for extremely delay-sensitive applications. This involves that the service should be deployed in the nearest EN, to guarantee the desired ultra-short latency.

Let $h_u(i)$ denote the probability of handovers to a near edge node EN_i for the u-th user. Indeed, for each user a mobility profile could be defined, to estimate its movements with good precision. Furthermore, let us define a matrix $P(u, n)$, which maps the user to the ENs according to the deployment of the relevant service instances. In this phase, we could assume that primary service instance and relevant replicas are equivalent.

The problem of managing replication management, such that both (i) the QoE degradation, i.e., the probability of reactive migrations, and (ii) the cost of replicas deployment are minimized, can be formulated according to the following linear program model:

$$\begin{aligned} \text{minimize} \quad & \sum_{u\in\mathcal{U}} \sum_{i\in\mathcal{N}} h_u(i)(1 - P(u, i)) \\ \text{minimize} \quad & \sum_{u\in\mathcal{U}} \sum_{i\in\mathcal{N}} P(u, i) \\ \text{s.t.} \quad & \forall u \in \mathcal{U}, \forall i \in \mathcal{N}, P(u, i) \in \{0, 1\} \\ & \forall u \in \mathcal{U}, P(u, i_u) = 1 \\ & \forall i \in \mathcal{N}, \sum_{u\in\mathcal{U}} P(u, i) \le INSTANCES_{MAX} \end{aligned}$$

The first objective aims to minimize the probability of reactive migration between different ENs. This objective can be satisfied by deploying additional service replicas only in edge nodes where the user is likely going to move to. The second goal aims at reducing the number of cost related to proactive service migration, i.e., the number of service replicas. Furthermore, the following constraints should be respected:

1) Constraint 1 ensures that the matrix P is binary, $\forall u \in \mathcal{U}$.
2) Constraint 2 guarantees that the primary instance of each user's application is deployed in i_u, i.e., the edge node which the u-th user is connected to.
3) Constraint 3 accounts for the limited resources of ENs, which can host a maximum number of service instances equal to $INSTANCES_{MAX}$.

V. Optimization Models for Replication-based Service Migration

We present two optimization solutions in order to resolve the multi-objectives problem described in the previous Section. The result of both solutions is represented by the matrix P, which provides decisions about the service deployment. The first solution is designed to support the highest QoE for mobile 5G applications with strict latency requirements at the network edge. In particular, it aims to minimize the reactive migrations for applications deployed in MEC system by exploiting replication mechanism, while guaranteeing that the number of service replicas does not overcome a specific threshold for each user. The second solution is proposed to minimize the cost of proactive service migration, while supporting the desired user QoE.

A. Min-RM: Minimizing Reactive Migration with Service Replication

Let $\mathcal{RM}$ denote the function that we aim at optimizing to achieve the highest user QoE. This function $\mathcal{RM}$ can be formally defined as the maximum probability of reactive migrations that can be tolerated in the MEC system for ultra-short delay critical applications. Furthermore, let $REPLICAS_{MAX}$ define the maximum number of deployable replicas for each activated service, which directly relates to the cost of replication mechanism each user is willing to sustain. If no limitation is imposed in terms of replication cost, then $REPLICAS_{MAX}$ could be set to ∞. In this case, the optimal solution for the user u-th would converge to deploying a replica in all edges nearby the current serving cloudlets, if required resources are available in the MEC systems. The optimization model to minimize reactive migration by exploiting service replication can be formulated according to the following linear program:

$$\begin{aligned}
\text{minimize} \quad & \mathcal{RM} \\
\text{s.t.} \quad & \forall u \in \mathcal{U}, \forall i \in \mathcal{N}, P(u,i) \in \{0,1\} \\
& \forall u \in \mathcal{U}, P(u,i_u) = 1 \\
& \forall i \in \mathcal{N}, \sum_{u \in \mathcal{U}} P(u,i) \leq INSTANCES_{MAX} \\
& \forall u \in \mathcal{U}, \sum_{i \in \mathcal{N}} P(u,i) \leq REPLICAS_{MAX} \\
& \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{N}} h_u(i)(1 - P(u,i)) \leq \mathcal{RM}
\end{aligned}$$

B. Min-NSR: Minimizing Number of Service Replicas for Proactive Migration

Similar to the previous solution, let $\mathcal{NSR}$ denote the function we aim at optimizing to reduce the number of service replicas and, therefore, the overall cost of service replication. Formally, we define $\mathcal{NSR}$ the maximum number of service replicas which could be deployed for each user application in the system. On the other hand, we set the maximum threshold of reactive migrations TRM_{MAX}, which is allowed to guarantee the desired user QoE. The optimization model to minimize the cost of service replication can be formulated according to the following linear program:

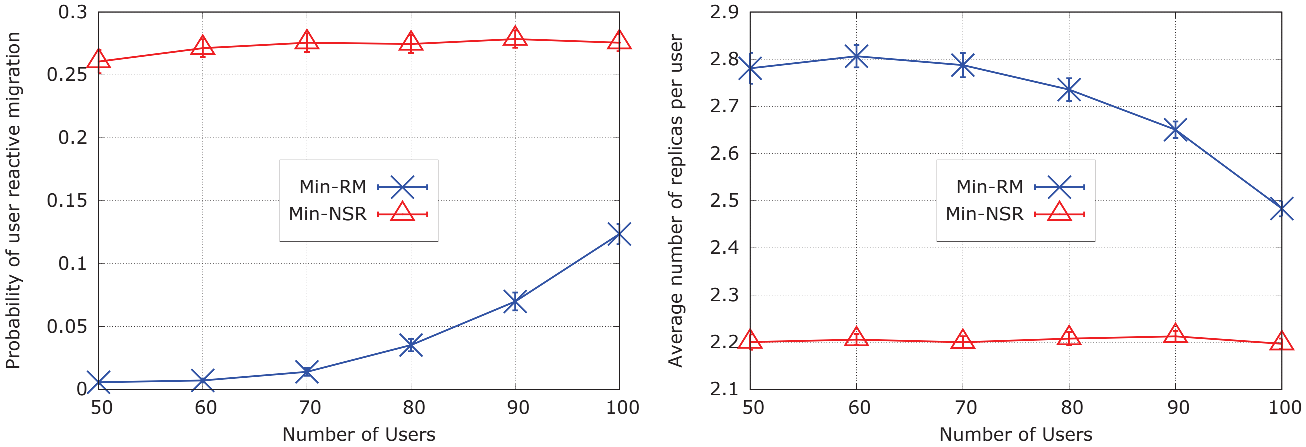
$$\begin{aligned}
\text{minimize} \quad & \mathcal{NSR} \\
\text{s.t.} \quad & \forall u \in \mathcal{U}, \forall i \in \mathcal{N}, P(u,i) \in \{0,1\} \\
& \forall u \in \mathcal{U}, P(u,i_u) = 1 \\
& \forall i \in \mathcal{N}, \sum_{u \in \mathcal{U}} P(u,i) \leq INSTANCES_{MAX} \\
& \forall u \in \mathcal{U}, \sum_{i \in \mathcal{N}} h_u(i)(1 - P(u,i)) \leq TRM_{MAX} \\
& \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{N}} P(u,i) \leq \mathcal{NSR}
\end{aligned}$$

VI. Performance Evaluation

To evaluate our proposed optimal solutions, we developed a simulator in Python, based on the Gurobi optimization framework [26]. The Gurobi Optimizer is an engine used by over 1500 companies and academic institutions to implement and solve complex mathematical models. The MEC simulation environment considers a mobile network composed of edge nodes, equi-spatially distributed in a grid topology over a small urban area of [1000, 1000] m. The edge nodes are interconnected through X2 LTE channels and can be represented by LTE eNodeB, whose radius coverage is around 250 meters for micro cells deployment. For our analysis we consider eNodeB enhanced with small-scale cloud capabilities and we consider a scenario where all the involved users require the same typology of service. This assumption is compliant with the emerging paradigm of network slicing [27], according to which definite amount of computing, storage, and networking resources are allocated to support the requirements of specific verticals in the edge-enhanced Telco network. In this context, we assume each EN can host a limited number of users' instances, e.g., 30, related to a particular vertical. We account for a varying number of users and, for each user, the MEC system is able to estimate the probability of handover in a near edge node, by using mobility prediction algorithms, such as [24], [23]. The handover probabilities to near edge nodes are dynamically stored for the i-th user in the relevant $h_u(i)$ matrix. In this way, the optimization solutions can analyze and decide the service replica deployment on-demand, according to the last registered user position and to the up-to-date handover probabilities. Furthermore, we assume that only one service is associated to each user. The proposed algorithms are evaluated in terms of the following two metrics:

- *Probability of user reactive migration*: it accounts for the QoE degradation, depending on the probability of reactive migration when a user moves to a nearby edge node.
- *Average number of replicas per user*: it is defined as the average number of service replicas deployed for each user service in the near edge nodes. This parameter allows for characterizing the cost of proactive replication-based migration.

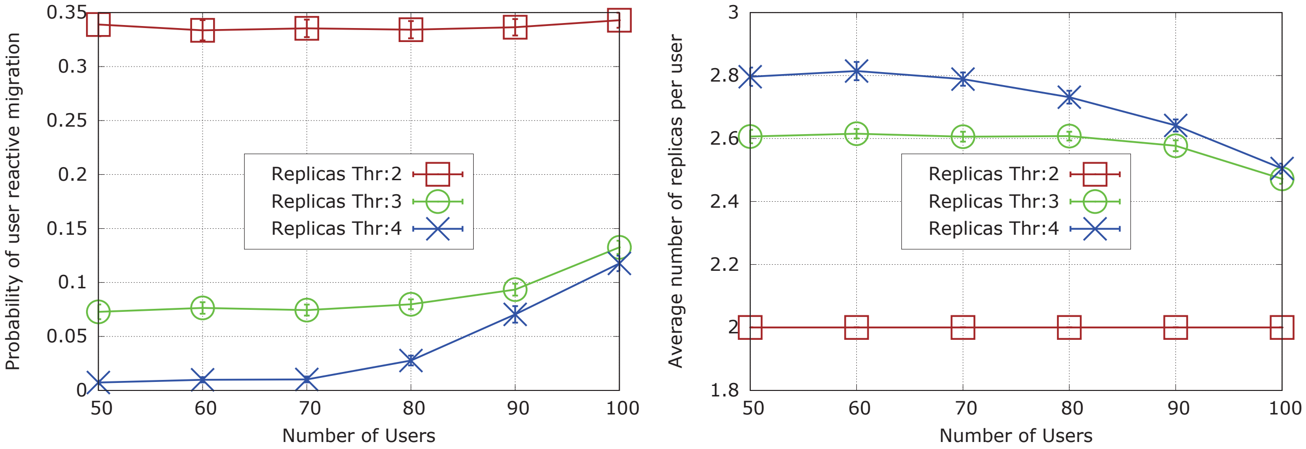
In the first analysis, we have compared the two proposed optimization solutions as a function of the number of users, as reported in Fig. 2. Each plotted point represents the average of 50 different runs of executions. The relevant curves are presented with 95 % confidence interval. In particular, for the *min-RM* solution, we have considered a maximum threshold of 4 service replicas for each user application, whereas we have set 0.5 as the upper-bound threshold of reactive migration for the *min-NSR* approach. Fig. 2(a) shows that *min-RM* outperforms *min-NSR* by guaranteeing a lower probability of reactive migration regardless the number of users. Instead, Fig. 2(b) illustrates that *min-NSR* allows for achieving a reduced cost of service replica deployment compared to the *min-RM* approach. The results obtained from the conducted simulations demonstrate the efficiency of each proposed solution in achieving its key design goals. Furthermore, we remind here that a higher probability of reactive migration implies a greater necessity to perform a relocation of service instance,



(a)

(b)

Fig. 2. Performance of service replication optimization solutions in terms of (a) probability of user reactive migrations and (b) average number of service replicas per user.



(a)

(b)

Fig. 3. Impact of different service replica thresholds on the performance of *min-RM* solution.

as a response of user movement. During this migration, the user can experience longer application response times, thus suffering from QoE degradation. Also, it is worth underlining that the probability of user reactive migration increases for *min-RM* solution by considering a higher number of users. This is due to the restricted capabilities of ENs, which limit the maximum number of deployed replicas instances per user.

In the second analysis, we evaluated the impact of different service replicas thresholds for the *min-RM* solution, by varying the number of users. In particular, we set the maximum number of service replicas to 2, 3, and 4. As shown in Fig. 3, the average probabilities of reactive migration take small values by increasing the relevant service replicas thresholds, since users can leverage more service replicas deployed in the nearby ENs. However, when the number of users with a higher threshold of service replicas increases, the edge systems is not able to allocate all the requested service replicas, due to the limited resources of ENs, and the relevant average number of deployed replicas decreases, as shown in Fig. 3(b). As a consequence, for replicas thresholds equal to 3 and 4, the resulting average probability of reactive migration increases when a higher number of users is considered.

VII. Conclusions and Open Research Areas

One key vision of the upcoming 5G network deals with the the support of ultra-short latency applications, by introducing service provisioning at the edge of the network. To guarantee the strict latency requirements, new solutions are required to cope with the user mobility in a distributed edge cloud environment. The use of proactive replication mechanism seems promising to avoid QoE degradation during service migration between different edge nodes. However, accounting for the limited resources of edge micro data-centers, appropriate optimization solutions must be developed to reduce the cost of service deployment, while guaranteeing the desired QoE.

In this paper, we have proposed two linear optimization solutions for replication-based service migration: the *min-RM* approach aims at minimizing the QoE degradation during user handover; *min-NSR* approach favors the reduction of service replication cost. Simulation results proved the efficiency of each solution in achieving its design goal and provides useful information for network and service orchestrators in next-generation 5G cloud-based networks.

Nonetheless, several research areas remain open and could be the subject of future studies, as highlighted in the following:

- *Path-oriented proactive migration*: The current optimization approaches are designed to be executed every time user moves to a different edge node. The proposed optimization approaches are based on Integer Linear Programming, whose solving times are suitable accounting for the low number of variables and a small-scale edge network. However, the execution of the optimization algorithms can be time consuming and suffer from scalability issues, especially for large network. A possible alternative can rely on the deployment of service replicas by leveraging accurate user path prediction from a source edge node to an estimated destination, thus requiring a reduced computation efforts for deciding the optimal service provisioning. Further studies may also be carried out to evaluate the sensitivity of the optimization approaches with respect to the accuracy of the user mobility prediction schemes.
- *Fair objectives trade-off*: As evaluated in this paper, probability of reactive migration and cost of service replication represent two conflicting objectives. To evaluate a fair trade-off between the analyzed goals, different models based on Game Theory may be further investigated.

Acknowledgment

This work was partially supported by the TAKE 5 project funded by the Finnish Funding Agency for Technology and Innovation (TEKES) and in part by the Finnish Ministry of Employment and the Economy. It is also partially supported by the European Unions Horizon 2020 research and innovation programme under the 5G!Pagoda project with grant agreement No. 7231722.

References

[1] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, 2009.

[2] "The tactile internet," ITU-T Technology Watch Report, Tech. Rep., 2014.

[3] ETSI ISG MEC, "Mobile-edge computing-introductory technical white paper," 2014.

[4] "OpenFog consortium," https://www.openfogconsortium.org/, accessed: October 2016.

[5] K. Ha, Y. Abe, Z. Chen, W. Hu, B. Amos, P. Pillai, and M. Satyanarayanan, "Adaptive vm handoff across cloudlets," Technical Report CMU-CS-15-113, CMU School of Computer Science, Tech. Rep., 2015.

[6] S. K. Bose, S. Brock, R. Skeoch, and S. Rao, "Cloudspider: Combining replication with scheduling for optimizing live migration of virtual machines across wide area networks," in *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*. IEEE, 2011, pp. 13–22.

[7] M. Kablan, B. Caldwell, R. Han, H. Jamjoom, and E. Keller, "Stateless network functions," in *Proceedings of the 2015 ACM SIGCOMM Workshop on Hot Topics in Middleboxes and Network Function Virtualization*. ACM, 2015, pp. 49–54.

[8] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "Ease: Epc as a service to ease mobile core network deployment over cloud," *Network, IEEE*, vol. 29, no. 2, pp. 78–88, 2015.

[9] H. Kang, M. Le, and S. Tao, "Container and microservice driven design for cloud infrastructure devops," in *2016 IEEE International Conference on Cloud Engineering (IC2E)*, April 2016, pp. 202–211.

[10] I. Farris, T. Taleb, A. Iera, and H. Flinck, "Lightweight Service Replication for Ultra-Short Latency Applications in Mobile Edge Networks," in *2017 IEEE International Conference on Communications (ICC)*, 2017.

[11] I. Farris, T. Taleb, H. Flinck, and A. Iera, "Providing Ultra-Short Latency to User-centric 5G Applications at the Mobile Network Edge," *Transactions on Emerging Telecommunications Technologies*, 2017.

[12] T. Taleb, S. Dutta, A. Ksentini, I. Muddesar, and H. Flinck, "Mobile edge computing potential in making cities smarter," *IEEE Communications Magazine*, 2017.

[13] R. W. Ahmad, A. Gani, S. H. A. Hamid, M. Shiraz, A. Yousafzai, and F. Xia, "A survey on virtual machine migration and server consolidation frameworks for cloud data centers," *Journal of Network and Computer Applications*, vol. 52, pp. 11–25, 2015.

[14] C. Colman-Meixner, C. Develder, M. Tornatore, and B. Mukherjee, "A survey on resiliency techniques in cloud computing infrastructures and applications," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[15] R. Morabito, J. Kjallman, and M. Komu, "Hypervisors vs. Lightweight Virtualization: a Performance Comparison," in *Cloud Engineering (IC2E), 2015 IEEE International Conference on*. IEEE, 2015, pp. 386–393.

[16] W. Li, A. Kanso, and A. Gherbi, "Leveraging linux containers to achieve high availability for cloud services," in *Cloud Engineering (IC2E), 2015 IEEE International Conference on*. IEEE, 2015, pp. 76–83.

[17] M. Rabinovich, Z. Xiao, and A. Aggarwal, "Computing on the edge: A platform for replicating internet applications," in *Web content caching and distribution*. Springer, 2004, pp. 57–77.

[18] P. A. Frangoudis, L. Yala, A. Ksentini, and T. Taleb, "An architecture for on-demand service deployment over a telco cdn," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[19] S. Dutta, T. Taleb, P. A. Frangoudis, and A. Ksentini, "On-the-fly qoe-aware transcoding in the mobile edge," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[20] A. Sgora and D. D. Vergados, "Handoff prioritization and decision schemes in wireless cellular networks: a survey," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 4, pp. 57–77, 2009.

[21] T. Taleb, "Toward carrier cloud: Potential, challenges, and solutions," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 80–91, June 2014.

[22] T. Taleb and A. Ksentini, "An analytical model for follow me cloud," in *2013 IEEE Global Communications Conference (GLOBECOM)*, Dec 2013, pp. 1291–1296.

[23] A. Nadembega, T. Taleb, and A. Hafid, "A destination prediction model based on historical data, contextual knowledge and spatial conceptual maps," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 1416–1420.

[24] G. Xu, S. Gao, M. Daneshmand, C. Wang, and Y. Liu, "A Survey for Mobility Big Data Analytics for Geolocation Prediction," *IEEE Wireless Communications*, 2016.

[25] A. Nadembega, A. Hafid, and T. Taleb, "An integrated predictive mobile-oriented bandwidth-reservation framework to support mobile multimedia streaming," *IEEE Transactions on Wireless Communications*, vol. 13, no. 12, pp. 6863–6875, Dec 2014.

[26] "Gurobi Optimizer - State of the Art Mathematical Programming Solver," http://www.gurobi.com/products/gurobi-optimizer, accessed: October 2016.

[27] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-end network slicing for 5g mobile networks," *Journal of Information Processing*, vol. 25, pp. 153–163, 2017.