The Location Problem for the Provisioning of Protected Slices in NFV-Based MEC Infrastructure

Hernani D. Chantre and Nelson Luis Saldanha da Fonseca

***Abstract*—The support of stringent requirements such as ultra-low latency and ultra-reliability of the forthcoming 5G services poses several challenges to telecommunications infrastructure providers. Network Function Virtualization, multi-access edge computing (MEC), and network slicing capabilities can help the support of such requirements. However, a trade-off between the cost of resource deployment and the support of service requirements needs to be taken into account in the design of NFV-based 5G networks. In this paper, we investigate the MEC location problem, which aims at selecting locations to place MECs hosting protected slices. We propose a MEC location problem enhanced with 1: 1 and $1 : N$ protection schemes for the provisioning of protected slices. In the 1: 1 scheme, protection is assured by reserving a backup slice for each tenant, whereas in the $1 : N$ scheme, a backup slice is shared among N tenants. The problem is modeled as a multi-criteria optimization problem and solved by the employment of a multi-objective evolutionary non-dominated sorting genetic algorithm. A comparison between the 1: 1 and $1 : N$ protection schemes is carried out in the context of 5G network slicing. Results show that the protection scheme 1: 1 can reduce the response time, at a higher deployment cost when compared to the $1 : N$ scheme.**

***Index Terms*—MEC location problem, protection schemes, network slicing, multi–access edge computing, 5G, NFV.**

I. Introduction

THE fifth-generation (5G) mobile telecommunications network supports diverse requirements of a wide range of services. 5G mobile networks encompass technologies such as a multi-access edge computing (MEC), network function virtualization (NFV), and network slicing through edge/cloud computing and virtualization/softwarization.

MECs deliver cloud computing services at the edge of a radio access network (RAN) [1], allowing the access to computational and storage resources close to the end-user device [1]–[4]. Instead of forwarding traffic to the mobile backhaul, customer requests are handled locally at the network edge, decreasing the load on the mobile backhaul and avoiding network bottlenecks [5]. Moreover, proximity to the end-user helps the support of stringent latency requirements of 5G services and use cases.

On the other hand, the NFV technology provides programmability to the management of the network infrastructure, enabling the decomposition of monolithic network functions into smaller Virtualized Network Functions (VNFs). NFV allows the virtualization of the core and RAN network functions (e.g., Mobility Management Entity, MME; Serving Gateway, SGW; Packet Data Network Gateway, PGW; Baseband Unit.), and enables the customization of network services. NFV brings flexibility and speed to the deployment of services. VNFs can be hosted on Virtual Machines (VM) or containers instantiated over central or distributed edge cloud computing systems [6]–[9].

Network slicing allows the allocation of segments of physical resources. A slice can be reserved for a VNF or to a chain of VNFs for service composition [10]. Therefore, network slicing plays a crucial role in support of a wide range of 5G applications and verticals (e.g., autonomous driving, tactile internet, augmented reality) with stringent and diverse requirements on top of a shared infrastructure. Slices, thus, must be adequately designed to support requirements such as reliability, latency, and availability.

The European Telecommunications Standards Institute (ETSI) introduced a MEC reference architecture, in which MECs are deployed as part of an NFV environment, and mobile edge applications are provisioned as VNFs [11]. A critical element in a MEC-based architecture is the MEC server, a general-purpose edge computing facility (node), that provides computing resources, storage capacity, connectivity, as well as radio and network information [12]. It can be deployed either at a base station (eNB) or at a multi-technology cell aggregation site (e.g., access points, switches, routers, and micro-datacenters).

The scenario considered in this paper includes an Infrastructure Provider (InP) owning a pool of geographically distributed MECs with a limited computational capacity [3]. The resources of these MECs can be sliced, and each slice allocated to a chain of VNFs implementing a request for service instantiation received by the InP. Slices are allocated on-demand and may have different capacities in terms of processing and memory. The goal of the InP is to devise

Manuscript received December 1, 2019; accepted March 2, 2020. Date of publication April 9, 2020; date of current version June 29, 2020. This work was supported in part by the PEC-PG-CAPES/CNPq Brazil, in part by the Sao Paulo Research Foundation (FAPESP) under Grant 15/24494-8, in part by the National Council of Scientific and Technological Development (CNPq), in part by the European Union's Horizon 2020 Project Federated Union of Telecommunications Research Facilities for an EU- Brazil Open Laboratory (FUTEBOL) under Agreement 688941, and in part by the Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTI) through Rede Nacional de Pesquisas (RNP). *(Corresponding author: Nelson Luis Saldanha da Fonseca.)*

Hernani D. Chantre is with the Institute of Computing, State University of Campinas, Campinas 13083-852, Brazil, and also with the Faculty of Science and Technology, University of Cape Verde, Praia 279, Cape Verde (e-mail: hernani@lrc.ic.unicamp.br).

Nelson Luis Saldanha da Fonseca is with the Institute of Computing, State University of Campinas, Campinas 13083-852, Brazil (e-mail: nfonseca@ic.unicamp.br).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2020.2986869

0733-8716 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

locations to MECs and hosted slices so that the requirements of services can be supported and cost minimized.

The provisioning of reliability in an NFV-based MEC deployment is critical, since a failure of a slice/MEC can cause a service outage, breaking the continuity of the hosted Service Function Chaining (SFC) [13], [14] [15]. To cope with failures, reliability can be assured by dynamic resource provisioning. Moreover, for latency-stringent services such as ultra-reliable and low-latency communications (uRLLC), the placement of a replica at a distant edge node can impact on the support of latency requirements [5], [16].

Protection schemes designate a backup (secondary) slice to serve the VNFs hosted by a slice (MEC) in failure to avoid service outage. In line with that, this paper evaluates traditional protection schemes such as 1: 1 (dedicated) and $1 : N$ (shared) to provide reliability to 5G services. In the 1: 1, a dedicated backup slice is reserved for each customer (demand point) while, in the $1 : N$ protection scheme, a backup slice (i.e., secondary) is shared among N demand points. In case of shared protection, protection service can be denied if the protecting slice is in use by another failed primary slice.

This paper proposes a formulation for the MEC location problem extended with protection schemes. The location problem is based on a generalized capacitated reliable facility location with a failure probability problem. The facilities are MECs that host slices in the form of a service chain of VNFs.

The contributions of this paper are:

i) The formulation of an extended MEC location problem with traditional protection schemes 1: 1 and $1 : N$;
ii) An evaluation of the trade-off between performance and deployment costs of 5G services;
iii) A comparison of traditional protection schemes 1: 1 and $1 : N$ for furnishing ultra-reliable services in 5G networks.

The rest of this paper is organized as follows. Section II discusses related work. Section III presents the statement of the problem, while the problem formulation is introduced in Section IV. Subsections IV-A and IV-B formulate the MEC location problem extended with the protection schemes 1: 1 and $1 : N$, respectively. Section V details the algorithm to solve the MEC location problem. Section VI presents numerical results, and Section VII concludes the paper.

II. Related Work

This section reviews the research on the MEC location problem for slice provisioning. The need for elasticity, flexibility, and reduced operational cost of mobile networks has motivated a large number of investigations on VNF placement over the cloud [17]–[20]. These papers have proposed algorithms for the placement of virtualized Evolved Packet Core elements (i.e., vEPC's, VNFs). In [21], Bagaa *et al.* proposed a VNF placement algorithm based on a coalition formation game, that derives the optimal number and locations of vEPC or VNFs elements over a federated cloud to host virtual instances of the vEPC elements. The proposed algorithm aims at ensuring QoS while reducing deployment costs. In [22], Carpio *et al.* studied the problem of VNF placement with replication to balance the network load and yet minimize server utilization. A Linear Programming (LP) model for the optimum placement of functions was proposed to minimize the link utilization and CPU usage. Laghrissi and Taleb [23] provided an extensive survey on the problem of VNF placement in clouds to accommodate 5G use cases (e.g., mobile broadband, Internet of Things, and autonomous driving) with different requirements such as mobility, latency, and reliability.

In [24], Dietrich *et al.* proposed a linear programming formulation for the computation of optimal virtualized EPC components as VNFs placement to achieve optimal load balancing, high request acceptance ratio, and high resource utilization. In [9], Taleb *et al.* proposed VNF placement algorithms for a carrier cloud to place P-GWs and S-GWs aiming at minimizing the length of the paths between users and their associated data anchor gateways.

New architectural models have been introduced to facilitate the provisioning of resources to the edge of the network [5], [25], [26]. In [27], Gouareb *et al.* studied the problem of VNF placement and routing across physical hosts to minimize overall queuing delay at the edge. In [28], Yang *et al.* presented a model to estimate the task completion delay and energy consumption of MECs. In [29], Bekkouche *et al.* investigated mobile edge computing in unmanned aerial vehicles (UAVs) and the effects of the latency and the reliability of the communication between the UAV Traffic Management (UTM) systems.

In [30], Laghrissi *et al.* presented a benchmark of VNF placement algorithms for a spatio-temporal model of mobile services over the distributed cloud at the edge to support ultra-low latencies services. The comparison was based on delay values, distance costs, and the frequency of VM overload periods. However, this work neither considered the capacity of the edge nodes, nor it provides alternatives in case of a node failure. Sarrigiannis *et al.* [31] presented a real-time allocation of VNFs scheme to a MEC-enabled 5G platform and cloud resources. They leveraged real-time services scale-out and scale-in features to handle the latency requirements of critical applications.

Daneshfar *et al.* [32] proposed an integer optimization model to address the problem of mapping services demands on infrastructure resources. The problem considered the randomness of resource availability in a MEC infrastructure. It aimed at minimizing the total cost of providing services while allocating demands to available resources. In [33], Jemma *et al.* considered the problem of VNF placement and provisioning over an edge cloud infrastructure. The authors took into account the QoS requirements as a multiple objective decision-making problem to maximize resource utilization, and reduce overloads. In [5], Yala *et al.* presented a placement scheme applicable to a MEC in an NFV environment tailored to uRLLC services. This work evaluated the trade-off between service access latency and availability. A genetic meta-heuristic algorithm was employed to solve the optimization problem.

Similar to our approach, [5], [33] addressed the problem of the VNF placement problem. The difference is that we formulated a multi-objective problem to activate MEC to furnish MEC slices for the allocation of different service

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

requests. The need for assuring the reliability of services that are geographically distributed in MECs introduces various service dependencies that prevent the satisfaction of Service Level Agreements (SLA) [26]. Different approaches to the VNF placement problem have considered reliability as a requirement [34]. However, these approaches have focused mainly on network failures [35].

Afolabi *et al.* [10] presented an implementation of an E2E network slice orchestration platform to deploy customized network slices according to the requirements of the services. The work in [36] investigated a robust VNF provisioning problem for minimizing the number of instantiated VNFs while maximizing the robustness of service. The work in [37] investigated end-to-end service reliability in Data Center Networks with the flow and SFCs parallelism and evaluated the number of backup VNFs required to protect parallelized SFC. In [38], the authors presented optimal strategies for the placement of edge resources in 5G networks and optimized the placement of primary and backup 5G user-plane functions (UPFs) at the edge. However, the proposed model did not consider the capacity limitation of the edge nodes, which may affect the solution, especially when dealing with stringent requirements of 5G services.

Chang and Wang [39] proposed two adaptive replication schemes to support mobile cloud applications for MECs. A responsive placement algorithm (RPA) and an increment placement algorithm (IPA) were proposed to reduce costs and increase revenue by calculating the number of replicas, which allows nearby MEC servers to process requests as well as to reduce the transmission latency. The solution showed that the RPA incurred a higher operational cost but provided high profit. RPA is suitable for applications with stringent delay requirements, whereas the IPA performs better for applications with delay-tolerant requirements. Qu *et al.* [40] studied a reliability-aware joint VNF chain placement and flow routing optimization in cloud datacenter networks. The problem was formulated as a complex ILP with reliability constraints. The authors explored the trade-off between reliability, bandwidth, and computing resources consumption of service chains.

In [41], Chang *et al.* studied the problem of VNF replica placement for cloud computing, considering the overhead of data consistency among replicas in different datacenters. A replica placement algorithm was introduced to improve QoS provisioning and service reliability. However, this work did not consider the capacity limitation of edge resources.

Duan *et al.* [42] presented a scheme for a resilient NFV system using a distributed actor model to provide lightweight failure resilience and high-performance flow migration. However, this work considered neither the capacities of the nodes nor their reliability. In [43], the authors proposed a heterogeneous backup deployment scheme to deal with the reliability problems faced by SFC. The authors argued that heterogeneous redundant backups lead to more significant gains in improving SFC reliability. The objective function minimized the link bandwidth consumption. However, this work did not consider resource utilization of the nodes.

Although previous approaches attempted to provide reliability and also reduce service response time, loss probability, and costs, they did not consider different protection schemes. The present paper proposes a multi-objective genetic algorithm for MEC location and slice provisioning to minimize the number of MEC deployed, the number of hosted slices, and the total service response time. We formulate an extend MEC location problem with 1: 1 and $1 : N$ protection schemes. Finally, we also demonstrate the advantages of selecting protection schemes to design a reliable NFV-based MEC compliant with the requirements of the 5G services.

Table I summarizes the most relevant papers related to VNF placement in MEC solutions per their target objective.

III. Statement of the Problem

This paper considers a 5G network with a distributed MEC infrastructure, owned by an Infrastructure provider (InP) having a pool of MECs geographically distributed [45]. MECs leverage virtualization, and provide slices to host VNFs chains. MECs have limited computational and storage capacities as well as bandwidth to handle the traffic generated by diverse service requests.

The InP receives requests to provide customized slices from application service providers (ASP), over-the-top (OTT) application providers, and vertical industries. Demand points request the InP slices to host the SCF of services. Demand points can also generate flows of packets to slices. For instance, a demand point may request a slice to implement the SCF of Broadcast/Multicast live streaming, e.g., a chain of vBM-SC, vMBMS-GW, vMME, vMCE, and also generate the stream to be broadcasted [46], [47].

Different services have specific resource requirements, impacting the number of slices needed and the amount of resources demanded.

The malfunction of a MEC implies service disruption or outages. Therefore, it is of paramount importance to devise a solution to map service requests onto MECs, considering protection schemes to accommodate data-sensitive services with minimum deployment cost.

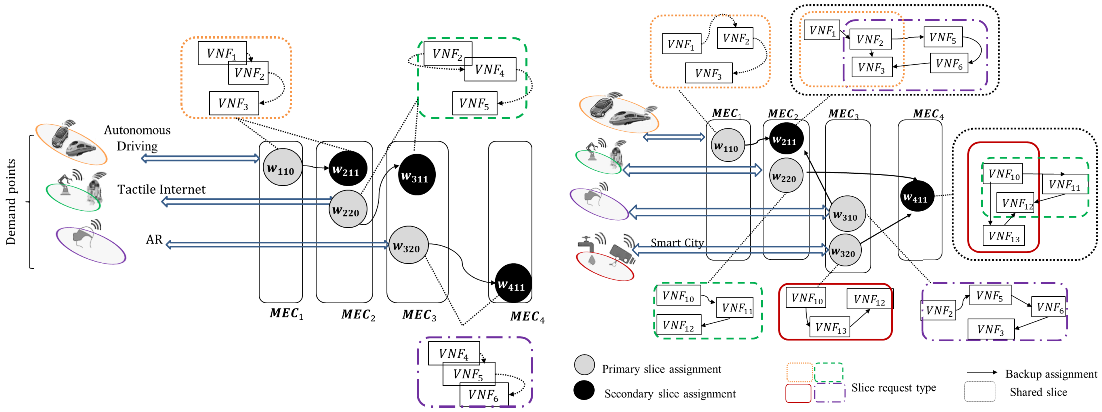
The formulation of the MEC location problem with traditional protection schemes such as 1: 1 and $1 : N$. A generic representation of MEC location solution with protection schemes is illustrated in Figure 1. Requests are assigned to a primary slice. In case of failure of a primary slice (i.e., first assignment) or its hosting MEC, the SCF is reassigned to a secondary slice hosted on a different MEC. Depending on the scheme implemented, the secondary slice can be shared (e.g., $1 : N$) or dedicated (e.g., 1: 1). Figure 1a illustrates the 1: 1 protection scheme, in which each demand point is assigned to a primary slice and a dedicated backup hosted in a different MEC. Fig.1a illustrates that the MEC_1 hosts one primary slice w_{110} for a given demand point and its dedicated backup slice w_{211} hosted on a distinct node MEC_2. Figure 1b shows the $1 : N$ protection scheme, in which reserved slices assigned as backup slices are shared among N demand points and illustrates the slice w_{211} in the MEC_2 node shared by two different demand points.

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARISON OF ASPECTS COVERED BY RELATED PAPERS

Approach	MEC	QoS metrics	Loss probability	Reliability	Cost of provisioning	Protection schemes	Multi-objective
[22]	No	Yes	No	No	No	No	Yes
[23]	No	Yes	No	Yes	Yes	Yes	Yes
[36]	No	No	Yes	Yes	No	Yes	No
[21]	No	Yes	No	No	Yes	No	Yes
[26]	Yes	Yes	No	Yes	No	No	No
[32]	Yes	Yes	No	Yes	Yes	No	No
[24]	Yes	Yes	No	No	No	No	No
[30]	Yes	Yes	No	No	No	No	No
[27]	Yes	Yes	No	No	No	No	No
[33]	Yes	Yes	No	No	Yes	No	Yes
[31]	Yes	Yes	No	No	Yes	No	No
[5]	Yes	Yes	No	Yes	Yes	No	Yes
[9]	Yes	Yes	No	No	No	No	Yes
[29]	Yes	Yes	Yes	Yes	No	No	No
[39]	Yes	Yes	No	No	Yes	Yes	Yes
[40]	No	Yes	Yes	Yes	No	No	Yes
[41]	No	Yes	No	Yes	No	Yes	Yes
[42]	No	Yes	Yes	No	No	Yes	No
[43]	No	Yes	No	Yes	No	Yes	No
[37]	No	Yes	Yes	Yes	No	Yes	No
[44]	Yes	Yes	No	Yes	No	No	No
our work	Yes	Yes	Yes	Yes	Yes	Yes	Yes



(a) 1 : 1 dedicated scheme.

(b) 1 : N shared scheme

Fig. 1. A generic representation of MEC location solution with protection schemes.

The employment of different protection schemes represents a design trade-off. For instance, the 1: 1 scheme calls for higher cost since it requires exclusively dedicated backup slices, and, consequently, a higher number of backup slices, when compared to the 1 : N scheme. On the other hand, the 1 : N implies lower deployment costs as the secondary slice is shared among tenants, which can incur extra latency since the location of the backup slice may not be optimal to all the demand points sharing the backup slice.

In particular, the proposed formulation tackles how an InP should design its infrastructure to deploy an optimized, reliable edge network considering 5G networks KPIs. The solution to this problem requires the determination of the number of MECs and slices to be deployed.

The MEC location problem is a network design problem solved off-line, which is typically used to decide on the design of a network to be deployed. Implementation details such as the specification of an agent to accept service requests and the migration of SCF are out of the scope of the present paper.

IV. Problem Formulation

This section introduces the formulation of a MEC location problem, which includes slice protection. The notation used in the formulation is summarized in Table II. Let $G = (U \cup V,\ E)$ be a bipartite graph in which U denotes a set of potential locations where a provider can activate MECs and V defines the set of demand points (tenants). The network link between

TABLE II

NOTATION USED IN THE MEC LOCATION PROBLEM FORMULATION

Symbol	Description
U	set of MECs
V	Set of demand points (tenants)
d_v	Size of requests of a demand point $v \in V$ in MB
K	Level of protection assignment of a slice (k=0 - primary, k=1 - secondary, ...)
B_u	Bandwidth capacity of a MEC u in Mbps
$b_{u\,v}$	Bandwidth between demand point $v \in V$ and MEC $u \in U$ in Mbps
Ψ	CPU capacity of the MECs in MIPS
Φ	RAM capacity of the MECs in GB
$\psi_{u\,i}$	CPU capacity of the i-th slice hosted in MEC u in MIPS
$\phi_{u\,i}$	RAM capacity of the i-th slice hosted in MEC u in GB
C_u	Maximum number of slices to be hosted in a MEC $u \in U$
τ_v	Processing (CPU) requirement of a demand point v in MIPS
σ_v	Memory (RAM) requirement of a demand point v in GB
$l_{u\,v\,i}$	Latency incurred by demand $v \in V$ using an i-th slice located at MEC $u \in U$ in ms
L_v	Maximum allowed latency for a demand point $v \in V$ in ms
r_u	Reliability of a MEC at the location $u \in U$
R_v	Required reliability level of a demand point $v \in V$
q	Failure probability of a MEC

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

the MEC and demand points are defined by $E \subseteq U \times V$. MECs $u \in U$ hosts a number of slices to allocate the service requests from the demand points $v \in V$. MECs are homogeneous and have CPU (Ψ) and RAM (Φ) capacities. The maximum number of slices to be hosted on a MEC is defined by the minimum between i) the ratio between the MEC CPU capacity and the minimum CPU demand of the demand points ($\min_v\{\tau_v\}$), and ii) the ratio between the MEC RAM capacity and the minimum RAM capacity of any demand point in ($\min_v\{\sigma_v\}$), i.e, $C_u = min(\frac{\Psi}{\min_v\{\tau_v\}}, \frac{\Phi}{\min_v\{\sigma_v\}})$.

We assume that MECs can fail due to different reasons such as unexpected restart/shutdown and power outages. Each MEC $u \in U$ is associated with a reliability value, r_u, which defines its probability of being available. The hosted slices inherit the reliability of their hosting MECs. The slices hosting a demand point are categorized either as primary or as secondary (i.e., $k = 0$ primary slices, $k = 1$ secondary slices), depending on its role in protecting a slice. In case a primary slice fails, a slice assigned as a backup (i.e. secondary slice) hosts the demands of the failed slice.

The goal is to find optimal locations to MECs hosting slices used by VNFs that compose a service requested. We address this objective by formulating a MEC location problem extended with protection schemes, which finds:

- The optimal number of slices to be hosted on the selected MEC;
- The optimal number of MECs to activate;
- The optimal assignment of demand points to the slices which incurs in minimum response time for the services;

The solution for the MEC location problem is given by a multi-objective criteria formulation which employs the following decision variables:

- $y_u \in \{0, 1\}$- the value 1 indicates an active MEC u is active (i.e., to host the serving slices).
- $w_{uik} \in \{0, 1\}$- the value 1 indicates that the i-th slice used as k assignment ($k = 0$ primary slice, $k = 1$ secondary slice) is hosted on MEC u.
- $x_{uvik} \in \{0, 1\}$ - the value 1 indicates that the i-th slice used as k assignment (i.e., $k = 0$ primary slice, $k = 1$ secondary slice) hosted on MEC $u \in U$ serves a demand point $v \in V$.

The multi-objective formulation has three objective functions:

$$Min \ \sum_{u\in U}\sum_{i=1}^{C_u}\sum_{k\in K} w_{uik} \tag{1}$$

$$Min \ \sum_{u\in U} y_u \tag{2}$$

$$Min \ \sum_{v\in V}\Big[\sum_{u\in U\setminus\{m\}}\sum_{i=1}^{C_u} d_v(b_{uv})^{-1}x_{uvi0}(1-q) + \sum_{m\in U\setminus\{u\}}\sum_{j=1}^{C_m} d_v(b_{mv})^{-1}x_{mvj1}q\Big] \tag{3}$$

The constraints of the problem are the following:

$$\sum_{i=1}^{C_u}(x_{uvi0} + x_{uvi1}) = 1 \quad \forall v \in V, \ u \in U \tag{4}$$

$$\sum_{u\in U}\sum_{i=1}^{C_u} x_{uvi0} = 1 \quad \forall v \in V \tag{5}$$

$$\sum_{u\in U}\sum_{i=1}^{C_u} x_{uvi1} = 1 \quad \forall v \in V \tag{6}$$

$$x_{uvik} \le w_{uik} \quad \forall u \in U, v \in V, k \in K, i \in 1..C_u \tag{7}$$

$$w_{uik} \le y_u \quad \forall u \in U, \ k \in K, \ i \in 1..C_u \tag{8}$$

$$x_{uvik} \le y_u \quad \forall u \in U, \ v \in V, \ k \in K, \ i \in 1..C_u \tag{9}$$

$$\sum_{i=1}^{C_u} \psi_{ui} \le \Psi \quad \forall u \in U \tag{10}$$

$$\sum_{i=1}^{C_u} \phi_{ui} \le \Phi \quad \forall u \in U \tag{11}$$

$$\sum_{v\in V}\sum_{i=1}^{C_u}\sum_{k\in K} b_{uv}x_{uvik} \le B_u \quad \forall u \in U \tag{12}$$

$$1 - \prod_{u\in U}[(1-r_u)^{x_{uvi0}}]\prod_{m\in U}[(1-r_m)^{x_{mvi1}}] \ge R_v \quad \forall v \in V, i \in 1..C_u, m \neq u. \tag{13}$$

$$\sum_{u\in U}\sum_{k\in K}\sum_{i=1}^{C_u} l_{uvi}x_{uvik} \le L_v \quad \forall v \in V \tag{14}$$

The objective functions defined by Equations (1), (2), (3) aim at minimizing the total number of slices, the number of active MECs and the total response time to serve a demand point, respectively. The trade-off expressed by these functions can be interpreted as: the higher the number of backups slices, the greater is the service reliability and the lower are the response times, but the greater is the cost incurred.

Constraint (4) indicates that a primary and a secondary slice of a demand point v" should not be hosted in the same MEC. Constraints (5) and (6) ensure that a given demand point v is allocated exactly to one primary slice and one secondary slice, respectively. Constraint (7) indicates that a demand point cannot be allocated on an unassigned slice. Constraint (8) indicates that a slice can not be hosted on an inactive MEC. Constraint (9) indicates that a demand point cannot be allocated on an inactive MEC.

Constraints (10) and (11) ensure that the total sum of CPU and RAM capacities of the slices hosted on a MEC $u \in U$ cannot exceed the capacity of the hosting MEC, respectively. Constraint (12) limits the bandwidth demand on MEC u can support. The reliability constraint (13) assures that the overall reliability level achieved with the implemented protection scheme should be greater than or equal to R_v, which is the expected reliability level of a demand point $v \in V$. The left-hand side of expression (13) checks if the probability of either the primary slice or the secondary slice hosted on different MECs is available.

Constraint (14) indicates that the total latency should be less than the maximum latency allowed for a demand point. L_v represents the maximum latency of a demand $v \in V$. The latency value is the processing time a demand point incurs on its i-th assigned slice hosted on MEC u (l_{uvi}), which is a

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

function of the CPU demand τ_v by the CPU capacity of the slice in use, $l_{uvi} = \frac{\tau_v}{\psi_{ui}}$.

The protection schemes 1: 1 and $1 : N$ are considered in the MEC location problem to furnish different protect services in case of MEC failure. They are presented in Subsection IV-A and IV-B.

A. 1: 1 *Protection Scheme*

The formulation of the MEC location problem presented in Section IV is extended to furnish a 1: 1 dedicated protection scheme to mitigate the impact of failures of the MEC and its hosted slices on service provisioning. The allocated slices to a demand point are categorized as primary or as secondary. In the event of a failure, the demand points are reassigned to a secondary slice (i.e., dedicated backup slice) hosted on a different MEC. Both primary and secondary slices are allocated to only one demand point; the secondary slice serves as a dedicated backup facility.

The objectives of the MEC location problem extended with 1: 1 scheme follows the objective functions (1), (2), (3).

The constraints (4), (5), (6), (7), (8), (12), (13) and (14) and those below defines the MEC location problem with 1: 1 scheme:

$$\sum_{v \in V} x_{uvik} \leq 1 \quad \forall u \in U, k \in K, i \in 1 \ldots C_u \tag{15}$$

$$\psi_{ui} \geq \tau_v x_{uvik} \quad \forall u \in U,\ v \in V,\ k \in K,\ i \in 1 \ldots C_u \tag{16}$$

$$\phi_{ui} \geq \sigma_v x_{uvik} \quad \forall u \in U,\ v \in V,\ k \in K,\ i \in 1 \ldots C_u \tag{17}$$

Constraint (15) indicates that either the primary or the secondary slice host only one demand point. Constraints (16) and (17) state that the CPU and RAM capacities of a slice assigned as primary or secondary slice should be greater than or equal to that requested by a demand point assigned to it.

B. $1 : N$ *Protection Scheme*

The formulation of the MEC location problem is extended to furnish a $1 : N$ shared protection scheme to mitigate the impact of failures of MECs and hosted slices on service provisioning. In the event of failure, the demand points are reassigned to a shared secondary slice hosted on a different MEC. The primary slices host exactly one demand point, and the secondary slices are shared by $N < |V|$ demand points.

The objectives of the MEC location problem extended with $1 : N$ scheme includes the objective functions (1), (2), (3).

The constraints (4), (5), (6), (7), (8), (12), (13) and (14) and those presented below define the MEC location problem with $1 : N$ protection scheme:

$$\sum_{v \in V} x_{uvi0} = 1 \quad \forall u \in U,\ i \in 1..C_u \tag{18}$$

$$\sum_{v \in V} x_{uvi1} \leq N \quad \forall u \in U,\ i \in 1..C_u \tag{19}$$

$$\psi_{ui} = \max_{v}\{\tau_v x_{uvi1}\} \quad \forall u \in U,\ v \in V,\ i \in 1..C_u \tag{20}$$

$$\phi_{ui} = \max_{v}\{\sigma_v x_{uvi1}\} \quad \forall u \in U,\ v \in V,\ i \in 1..C_u \tag{21}$$

$$\psi_{ui} \geq \tau_v x_{uvi0} \quad \forall u \in U,\ v \in V,\ i \in 1..C_u \tag{22}$$

$$\phi_{ui} \geq \sigma_v x_{uvi0} \quad \forall u \in U,\ v \in V,\ i \in 1..C_u \tag{23}$$

Constraint (18) indicates that a primary slice allocates exactly one demand point. Constraint (19) indicates that a secondary slice is shared by $N < |V|$ demand points. Constraint (20) and (21) indicate that the CPU and RAM capacities of a shared secondary slice is the maximum capacity demanded by the N demand points sharing that slice. Constraint (22) and (23) state that the CPU and RAM capacities of a primary slice should be greater than or equal to the CPU and RAM requirements of the demand point assigned to it.

V. Genetic Algorithm for MEC Location Problem

The MEC location problem, a facility location problem, is formulated as a constrained multi-objective optimization problem, which is a NP-hard problem. Thus, we employ a metaheuristics algorithm called the non-dominated Sorting Genetic Algorithm (NSGA-II) [48] to derive an optimal number of instantiated slices and MECs to support the latency and reliability requirements of 5G services.

The core part of the algorithm follows our previous work [49]. The intuition behind the algorithm is to find the set of solutions that are not dominated (Pareto front) by any other solutions, using ranking and crowding criteria. A solution is considered better than the others if it dominates them. Solution i dominates solution j if solution i is better or equal to solution j in terms of the criteria adopted, namely: response time, reliability, and the number of instantiated slices. Three different ranking techniques are used to find a non-dominated solution. The first raking uses only the objective function values. The second raking considers only the constraint violation values of all the constraints, and the third one combines the objective functions and constraint-violation values. At the end of the ranking process, the solutions with the best ranking are chosen.

The NSGA-II algorithm is a population-based algorithm based on the theory of natural selection and evolution of individuals. Individuals are potential solutions represented by chromosomes composed of genes (i.e., sub-chromosomes). Each gene encodes the number of slices a given MEC must host. The chromosome is translated into a solution to the problem.

The algorithm simulates the creation of the number of slices at each selected MEC location according to the service requests. First, a population set of individuals are randomly generated, selecting locations to activate MECs, computing the number of slices to be hosted at each MEC. Then a crossover operation is executed by randomly selecting two individuals from the population and produce offspring in such a way that the children inherit as much as possible of useful information from the two individuals. The mutation operation is then applied to each gene with a mutation probability. The mutated genes generate a new value that tries to produce a new population, emulating the creation of the number of slices

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

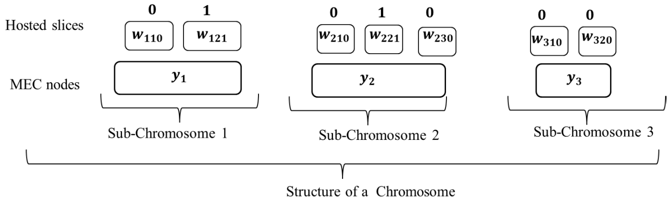


Fig. 2. Genetic encoding scheme for MEC location.

in every active MEC, to find a better solution for the MEC location problem. If the individual created is not valid, i.e., the algorithm cannot derive the number of slices to be hosted on an active MEC such that the latency and reliability requirements are fulfilled. It then discards the chromosome and creates a new individual using the crossover and mutation operation. Fig. 2 illustrates an example of genetic encoding for the MEC location problem with three genes or sub-chromosome (e.g., the number of active MECs), and the corresponding number of slices and their assignment level.

This process is repeated until achieving the desired population size. The algorithm assigns a fitness value that reflects the solution goodness concerning the optimization objectives, i.e., the number of MECs to be deployed, and the number of slices to be hosted in each MEC. As the MEC location problem is a minimization problem, the minimum fitness is ranked number one. At the end of the procedure, the result represents the configuration of an NFV-based MEC infrastructure with the optimal number of MECs to activate, the optimal number of slices to be hosted on the selected MECs, and optimal assignment of demand points to the slices having the minimum response time.

The computational complexity of the employed algorithm is $O(MpS^2)$, which is driven by the classification process of the non-dominated solutions set (Pareto front), where M is the number of objectives, and pS is the population size.

In our experiments, a simulated binary crossover operator for mating and polynomial mutation with the probabilities of 0.9 to perform crossover and (1/number of nodes) were employed, respectively. To evaluate the investigated NFV-based MEC location problem, a Java-based framework for multi-objective optimization, named JMetal Framework [50], was used.

VI. Numerical Results

This section presents the results obtained for different datasets as input to the location problem. We run a Multi-Objective Genetic Algorithm on a JMetal version 5.3 [50] and on a Debian GNU/Linux Squeeze, with two Intel Xeon (2.13GHz) with 4 cores each, and 78GB RAM. The population size for the algorithm was set to 100, the number of generations to 25, and a confidence level of 95% was used.

Following the work in [51], the infrastructure scenario for the MEC location problem is composed of MEC nodes, distributed in a grid topology over an ultra-dense 5G network in an area of 1000 x 1000 meters. The designed network infrastructure is composed of 10 to 60 MECs nodes. MECs nodes have 8 virtual CPUs, MIPS (4800), and 8GB of RAM capacities. The data rate supported by the MECs is 400 Mbps [52].

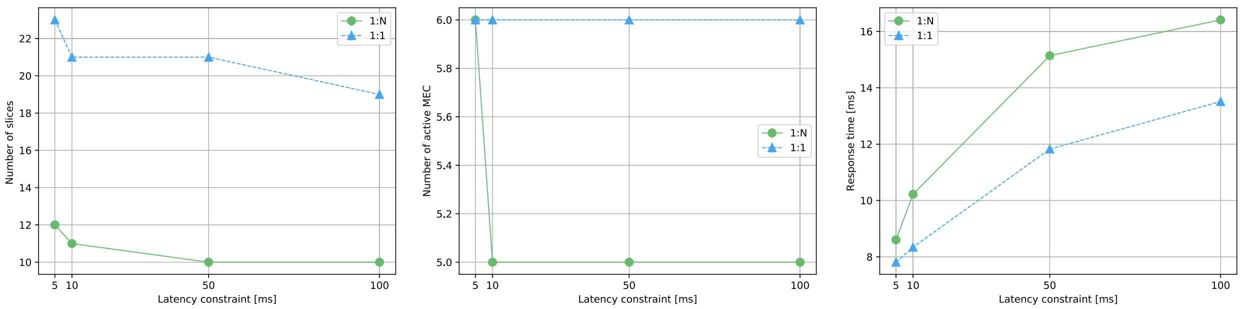
Demand points request service instantiations of different 5G verticals. These deployments are implemented as chains of VNFs. A VNF chain is hosted in a single slice. A requests for service instantiation is represented by the CPU, RAM, bandwidth demands of the chain VNFs implementng the service, as well as the latency and reliability requirements of the service. Latency values were considered in the interval [5,10,50,100] ms. Three class of reliability requirements are considered: high-level of reliability (99.999%), middle-level (99.9%), and low-level of reliability (9.0%). Demands of CPU and RAM are uniformly distributed in the range [0, 4] vCPUs and [0, 6] RAM, respectively. The data rate from the demand point modeled is uniformly distributed in the interval [100, 300] Mbps.

Figure 3 shows a comparison between the schemes 1:N and 1:1 for different latency constraints. We consider a scenario with the following latency constraint [5, 10, 50, 100] ms to represent different latency-sensitive service requests. We set $U = 10$ for the number of MECs, $V = 100$ the total of tenants, $N = 10$, which is the maximum number of total tenants sharing a secondary slice, $K = 2$, and $R_v = 99.999\%$ for highly reliable services.

The results depicted in Fig. 3a confirm that the employment of the scheme $1 : N$ leads to a substantial reduction in the number of hosted slices. Fig. 3a also confirms that as the latency constraints are relaxed, the number of hosted slices is reduced. Fig. 3b shows a slight reduction in the number of active MECs when the $1 : N$ scheme is employed. Fig. 3b shows that the effect of latency constraints in the total number of MECs to be activated is minimal for the $1 : N$, whereas for the 1: 1 scheme the number of MECs tends to be constant. Fig. 3c displays the service response time of the protection schemes as a function of the latency requirements. Fig. 3c shows a gain in the service response time of the scheme 1: 1 over the $1 : N$ scheme. For requests with latency requirement of 5ms, the service response time of the 1: 1 and $1 : N$ schemes were in the order of 6ms and 8ms, respectively. This figure shows an increase in the service response time for both schemes as the latency requirements are relaxed. The achieved response time was in the order of 16ms for less stringent demands with latency requirements of 100ms. This trend is confirmed by the results shown in Fig. 3b, in which the number of hosted slices tends to decrease when the latency requirements are relaxed.

We next evaluate the protection scheme taking into account the objectives of our schemes by comparing them against different reliability constraints [9.0%, 9.9%, 99.9%, 99.999%] to represent low-reliable, high-reliable and ultra-reliable service requests. Fig. 4 shows that the number of hosted slices and activated MECs is relatively lower for the scheme $1 : N$ when compared with that of the 1: 1 scheme. Results in Fig. 4c confirm that the 1: 1 produces lower response time than does the $1 : N$ scheme. We further observe that low-reliable service requests require fewer slices and MECs to be instantiated than do high reliable service requests, and, consequently, the

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

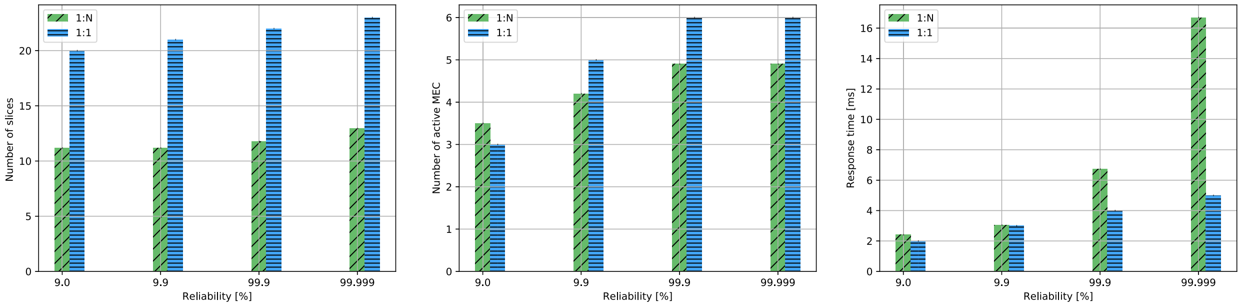


(a) The number of hosted slices.

(b) The number of activated MECs.

(c) Response time

Fig. 3. The number of hosted slices and activated MECs, as well as the response time as a function of latency constraints.

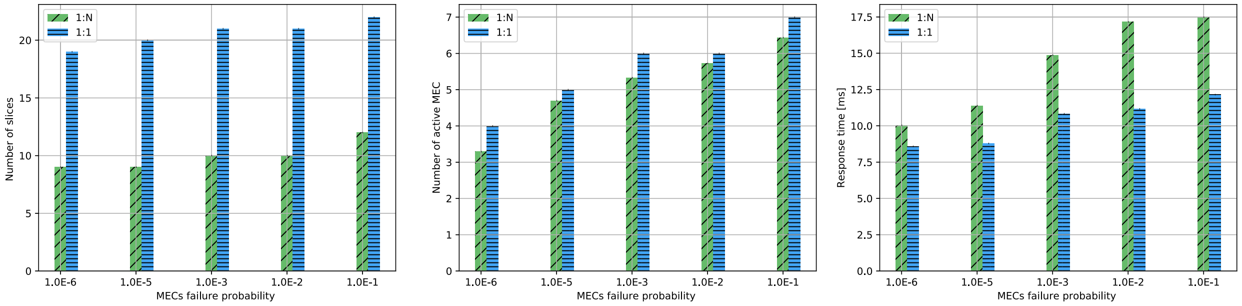


(a) The number of hosted slices.

(b) The number of activated MECs.

(c) Response time

Fig. 4. The number of hosted slices and activated MECs, as well as the response time as a function of reliability constraints.



(a) The number of hosted slices.

(b) The number of activated MECs.

(c) Response time

Fig. 5. The number of hosted slices and activated MECs, as well as response time as a function of MECs reliability r_u.

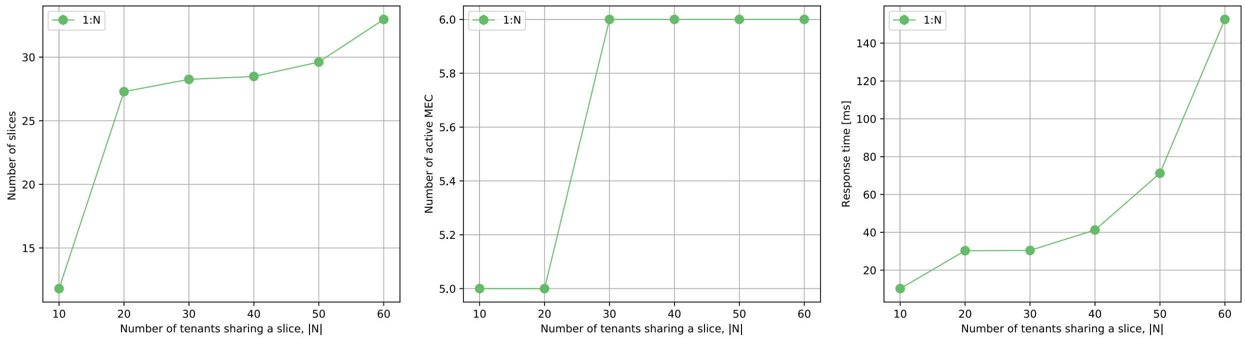
response service time is higher. Fig. 4a and 4b confirm that the results showed in Fig. 3a and 3b demonstrate that the scheme $1 : N$ requires lower deployment costs than does the 1: 1 scheme, since the number of hosted slices and MECs is smaller.

Fig. 5 evaluates the impact of different failure probability on the performance of the two protection schemes. As expected, the 1: 1 scheme demands a higher number of slices to be hosted than does the $1 : N$ scheme. On the other hand, the service response time achieved with that of the 1: 1 scheme is lower than that with the $1 : N$ scheme. This represents a trade-off in service response and the number of MECs to be deployed. Fig. 5a, 5b, and 5c demonstrate that the more reliable the MECs, the lower is the number of MECs and slices to be deployed.

We analyze the impact of the number of tenants N sharing a backup slice on the performance of the scheme $1 : N$ (Fig. 6). As expected, Fig. 6a shows that the number of slices increases as N increases. Moreover, Fig. 6b shows that, for $20 < N < 60$, the number of activated MECs remains constant around 6 MECs. Results in Fig. 6c confirm that the service response time grows with the number of tenants that shares a slice. This is related to the additional overhead incurred by the tenants on a shared slice.

In summary, results in this section show that furnishing a protection scheme implies a trade-off between deployment costs and service performance. Results evince that the $1 : N$ provides significant savings for an infrastructure provider while the 1: 1 scheme provides low service response times. Results demonstrate that from an InP perspective the $1 : N$

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.



(a) The number of hosted slices.

(b) The number of activated MECs.

(c) Response time

Fig. 6. The number of hosted slices and activated MECs, as well as the response time as a function of $|N|$.

scheme is preferable since it requires a smaller number of slices compared to the 1: 1 scheme, and, consequently, costs less. From a user perspective, the 1: 1 scheme is not only more reliable but can provide shorter response times since the slice is usually located at a place which minimizes the response time. In the $1 : N$ scheme, the backup slice is located at a place which does not necessarily minimize the response time of all the demand points hosted in the shared slice. On the other hand, the cost of reserving slices as dedicated backup is higher when compared to the cost of the shared scheme.

VII. Conclusion

In this paper, the problem of MEC location extended with protection schemes was investigated. The problem tackles how a provider should design its NFV-based MEC infrastructure to allocate service requests compliant with the requirements of 5G use cases. We studied the MEC location problem, furnishing traditional $1 : N$ and 1: 1 protection schemes. A multi-objective genetic algorithm optimization formulation was employed to derive a solution to the MEC location problem. Results indicate that there is a trade-off in deployment costs and service provisioning.

The $1 : N$ protection schemes requires a lower number of hosted slices and MECs, but the service response time can be affected. On the other hand, the 1: 1 scheme provides shorter latencies, but the deployment cost is higher than that of the $1 : N$ scheme as the number of hosting slices and MECs is larger.

References

[1] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[2] B. Han, S. Wong, C. Mannweiler, M. R. Crippa, and H. D. Schotten, "Context-awareness enhances 5G multi-access edge computing reliability," *IEEE Access*, vol. 7, pp. 21290–21299, 2019.

[3] R. Riggio, S. N. Khan, T. Subramanya, I. G. B. Yahia, and D. Lopez, "LightMANO: Converging NFV and SDN at the edges of the network," in *Proc. NOMS - IEEE/IFIP Netw. Oper. Manage. Symp.*, Apr. 2018, pp. 1–9.

[4] P.-V. Mekikis *et al.*, "NFV-enabled experimental platform for 5G tactile Internet support in industrial environments," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1895–1903, Mar. 2020.

[5] L. Yala, P. A. Frangoudis, and A. Ksentini, "Latency and availability driven VNF placement in a MEC-NFV environment," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[6] A. Ksentini, M. Bagaa, and T. Taleb, "On using SDN in 5G: The controller placement problem," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[7] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Combined virtual mobile core network function placement and topology optimization with latency bounds," in *Proc. 4th Eur. Workshop Softw. Defined Netw.*, Sep. 2015, pp. 97–102.

[8] L. Ma, X. Wen, L. Wang, Z. Lu, and R. Knopp, "An SDN/NFV based framework for management and deployment of service based 5G core network," *China Commun.*, vol. 15, no. 10, pp. 86–98, Oct. 2018.

[9] T. Taleb, M. Bagaa, and A. Ksentini, "User mobility-aware virtual network function placement for virtual 5G network infrastructure," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3879–3884.

[10] I. Afolabi, T. Taleb, P. A. Frangoudis, M. Bagaa, and A. Ksentini, "Network slicing-based customization of 5G mobile services," *IEEE Netw.*, vol. 33, no. 5, pp. 134–141, Sep. 2019.

[11] *Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NVF Environment*, document GR 017 V1.1.1, ETSI, 2018.

[12] F. Giust *et al.*, "MEC deployments in 4G and evolution towards 5G," *ETSI White Paper*, vol. 24, pp. 1–24, Feb. 2018.

[13] N. Shahriar, R. Ahmed, A. Khan, S. R. Chowdhury, R. Boutaba, and J. Mitra, "ReNoVatE: Recovery from node failure in virtual network embedding," in *Proc. 12th Int. Conf. Netw. Service Manage. (CNSM)*, Oct. 2016, pp. 19–27.

[14] *NFV; Reliability; Reports on Models and Features for End-to-End Reliability*, document GS NFV-REL 003 V1.1.1, ETSI, Apr. 2016.

[15] A. Zhou *et al.*, "Cloud service reliability enhancement via virtual machine placement optimization," *IEEE Trans. Services Comput.*, vol. 10, no. 6, pp. 902–913, Nov. 2017.

[16] W. Guan, X. Wen, L. Wang, Z. Lu, and Y. Shen, "A service-oriented deployment policy of end-to-end network slicing based on complex network theory," *IEEE Access*, vol. 6, pp. 19691–19701, 2018.

[17] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and S. Davy, "Design and evaluation of algorithms for mapping and scheduling of virtual network functions," in *Proc. 1st IEEE Conf. Netw. Softwarization (NetSoft)*, Apr. 2015, pp. 1–9.

[18] M. Bagaa, T. Taleb, A. Laghrissi, A. Ksentini, and H. Flinck, "Coalitional game for the creation of efficient virtual core network slices in 5G mobile systems," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 469–484, Mar. 2018.

[19] A. Laghrissi, T. Taleb, and M. Bagaa, "Conformal mapping for optimal network slice planning based on canonical domains," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 519–528, Mar. 2018.

[20] I. Benkacem, T. Taleb, M. Bagaa, and H. Flinck, "Optimal VNFs placement in CDN slicing over multi-cloud environment," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 616–627, Mar. 2018.

[21] M. Bagaa, T. Taleb, A. Laghrissi, and A. Ksentini, "Efficient virtual evolved packet core deployment across multiple cloud domains," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.

[22] F. Carpio, W. Bziuk, and A. Jukan, "Replication of virtual network functions: Optimizing link utilization and resource costs," in *Proc. 40th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2017, pp. 521–526.

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.

[23] A. Laghrissi and T. Taleb, "A survey on the placement of virtual resources and virtual network functions," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1409–1434, 2nd Quart., 2019.

[24] D. Dietrich, C. Papagianni, P. Papadimitriou, and J. S. Baras, "Network function placement on virtualized cellular cores," in *Proc. 9th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2017, pp. 259–266.

[25] P. Bellavista, L. Foschini, and D. Scotece, "Converging mobile edge computing, fog computing, and IoT quality requirements," in *Proc. IEEE 5th Int. Conf. Future Internet Things Cloud (FiCloud)*, Aug. 2017, pp. 313–320.

[26] S. N. Shirazi, A. Gouglidis, A. Farshad, and D. Hutchison, "The extended cloud: Review and analysis of mobile edge computing and fog from a security and resilience perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2586–2595, Nov. 2017.

[27] R. Gouareb, V. Friderikos, and A.-H. Aghvami, "Virtual network functions routing and placement for edge cloud latency minimization," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2346–2357, Oct. 2018.

[28] S. Yang, F. Li, M. Shen, X. Chen, X. Fu, and Y. Wang, "Cloudlet placement and task allocation in mobile edge computing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5853–5863, Jun. 2019.

[29] O. Bekkouche, T. Taleb, and M. Bagaa, "UAVs traffic control based on multi-access edge computing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[30] A. Laghrissi, T. Taleb, M. Bagaa, and H. Flinck, "Towards edge slicing: VNF placement algorithms for a dynamic & realistic edge cloud environment," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[31] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P. Mekikis, A. Antonopoulos, and C. Verikoukis, "Online VNF lifecycle management in a MEC-enabled 5G IoT architecture," *IEEE Internet Things J.*, early access, Oct. 1, 2019, doi: 10.1109/JIOT.2019.2944695.

[32] N. Daneshfar, N. Pappas, V. Polishchuk, and V. Angelakis, "Service allocation in a mobile fog infrastructure under availability and QoS constraints," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[33] F. Ben Jemaa, G. Pujolle, and M. Pariente, "QoS-aware VNF placement optimization in edge-central carrier cloud architecture," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–7.

[34] H. Chantre and N. L. S. D. Fonseca, "Reliable broadcasting in 5G NFV-based networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 218–224, Mar. 2018.

[35] H. D. Chantre and N. L. S. da Fonseca, "Redundant placement of virtualized network functions for LTE evolved multimedia broadcast multicast services," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[36] T. Lin and Z. Zhou, "Robust virtual network function provisioning under random failures on network function enabled nodes," in *Proc. 10th Int. Workshop Resilient Netw. Design Model. (RNDM)*, Aug. 2018, pp. 1–7.

[37] A. Engelmann and A. Jukan, "A reliability study of parallelized VNF chaining," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[38] I. Leyva-Pupo, A. Santoyo-González, and C. Cervelló-Pastor, "A framework for the joint placement of edge service infrastructure and user plane functions for 5G," *Sensors*, vol. 19, no. 18, p. 3975, 2019.

[39] W.-C. Chang and P.-C. Wang, "Adaptive replication for mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2422–2432, Nov. 2018.

[40] L. Qu, M. Khabbaz, and C. Assi, "Reliability-aware service chaining in carrier-grade softwarized networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 558–573, Mar. 2018.

[41] W.-C. Chang and P.-C. Wang, "Write-aware replica placement for cloud computing," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 656–667, Mar. 2019.

[42] J. Duan, X. Yi, S. Zhao, C. Wu, H. Cui, and F. Le, "NFVactor: A resilient NFV system using the distributed actor model," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 586–599, Mar. 2019.

[43] J. Xie, P. Yi, Z. Zhang, C. Zhang, and Y. Gu, "A service function chain deployment scheme based on heterogeneous backup," in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2018, pp. 1096–1103.

[44] L. Qu, C. Assi, K. Shaban, and M. J. Khabbaz, "A reliability-aware network service chain provisioning with delay guarantees in NFV-enabled enterprise datacenter networks," *IEEE Trans. Netw. Service Manage.*, vol. 14, no. 3, pp. 554–568, Sep. 2017.

[45] V. Scoca, A. Aral, I. Brandic, R. De Nicola, and R. B. Uriarte, "Scheduling latency-sensitive applications in edge computing," in *Proc. 8th Int. Conf. Cloud Comput. Services Sci.*, 2018, pp. 158–168.

[46] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: Overview and rel-11 enhancements," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[47] T. Taleb and A. Ksentini, "Gateway relocation avoidance-aware network function placement in carrier cloud," in *Proc. 16th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, 2013, pp. 341–346.

[48] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[49] H. D. Chantre and N. L. S. da Fonseca, "Multi-objective optimization for edge device placement and reliable broadcasting in 5G NFV-based small cell networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2304–2317, Oct. 2018.

[50] A. J. Nebro, J. J. Durillo, and M. Vergne, "Redesigning the jMetal multi-objective optimization framework," in *Proc. Companion Publication Genetic Evol. Comput. Conf. Companion (GECCO)*, 2015, pp. 1093–1100.

[51] I. Farris, T. Taleb, M. Bagaa, and H. Flick, "Optimizing service replication for mobile delay-sensitive applications in 5G edge network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[52] V. Scoca, A. Aral, I. Brandic, R. De Nicola, and R. B. Uriarte, "Scheduling latency-sensitive applications in edge computing," in *Proc. 8th Int. Conf. Cloud Comput. Services Sci.*, vol. 1, 2018, pp. 158–168.

Hernani D. Chantre received the B.S. degree in applied mathematics and informatics from RUDN, Russia, in 2006, and the M.Sc. degree in computer science from Bridgewater State University, MA, USA, in 2010. He is currently pursuing the Ph.D. degree in computer science with the State University of Campinas, Brazil. He is also an Assistant Graduate Professor at the University of Cape Verde. His research interests include network function virtualization and network-based cloud computing.

Nelson Luis Saldanha da Fonseca received the Ph.D. degree in computer engineering from the University of Southern California, Los Angeles, CA, USA, in 1994. He is currently a Full Professor at the Institute of Computing, State University of Campinas, Campinas, Brazil. He has authored or coauthored over 400 articles and has supervised over 60 graduate students. He was a recipient of the 2012 IEEE Communications Society (ComSoc) Joseph LoCicero Award for Exemplary Service to Publications, the Medal of the Chancellor of the University of Pisa, in 2007, and the Elsevier Computer Network Journal Editor of Year 2001 Award. He is currently the Vice President Technical and Educational Activities of the IEEE Communications Society (ComSoc). He served as the ComSoc Vice President Publications, the Vice President Member Relations, the Director of Conference Development, the Director of Latin America Region, and the Director of On-Line Services. He is the Past Editor-in-Chief of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He is a Senior Editor of the *IEEE Communications Magazine*, an Editorial Board Member of *Computer Networks* and *Peer-to-Peer Networking and Applications*.

Authorized licensed use limited to: IEEE Editors-in-Chief. Downloaded on August 04,2020 at 14:27:05 UTC from IEEE Xplore. Restrictions apply.