Design of 5G MEC-Based Networks With 1:N:K Protection Scheme

Hernani D. Chantre and Nelson L. S. da Fonseca, *Senior Member, IEEE*

***Abstract*—With the advent of 5G networks, telecommunications infrastructure providers (InP) have faced numerous challenges as they attempt to meet the stringent quality of service requirements. The placement of applications at the edge of the mobile network in Multi-access Edge Computing (MEC) and slicing techniques have provided powerful tools to enable networks to support these requirements. This paper studies the problem of locating MECs and slices in a 5G infrastructure protected by a $1 : N : K$ protection scheme. The aim is to support high reliability and low latency requirements at a minimum cost. A bi-objective non-linear formulation is proposed, and a solution is derived by employing the non-dominated sorting genetic algorithm (NSGA)-II. Results show that the enhanced $1 : N : K$ scheme is cost-effective. The proposal is evaluated on the basis of various levels of reliability, latency requirements, and probability of failure.**

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

I. Introduction

THE FIFTH-GENERATION cellular network (5G) technology provides an unprecedented quality of service (QoS) for end-users, including multi-Gbps data rates, high-reliability levels of five nines, and sub-millisecond latency. These requirements vary according to the specific 5G use cases, such as enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (uRLLC), and massive Machine Type Communications (mMTC) [1], [2]. By furnishing such an enhanced technology, 5G enables services in various vertical industries, such as healthcare and smart vehicles. However, the provisioning of services in 5G brings several challenges to telecommunications infrastructure providers (InP) in relation to Operational expenditure (OPEX), maintainability, and efficiency, especially for ultra-dense networks. Moreover, the demand for 5G services will require special attention in planning the network infrastructure to support stringent requirements of ultra-low latency simultaneously with high reliability [3].

Moreover, chains of Virtual Network Functions (VNFs) [1], [4], [5], [6], also known as Service Function Chains (SFC), can be instantiated in 5G multi-access edge computing (MEC) nodes, which can furnish computing resources, storage capacity, and connectivity, as well as radio and network information at the edge of the network [7], [8], [9], thus greatly facilitating the provisioning of low latency requirements [10], [11], [12].

These nodes can be deployed either at base stations (eNB) or in cells with multi-technology aggregation sites (e.g., access points, switches, routers, and micro data centers) in an ultra-dense 5G network. The resources of MEC nodes can be virtualized in slices, with a single MEC hosting several slices allocated to different end-users (demand point). These slices can be employed to execute VNFs or even SFCs. Moreover, isolation techniques guarantee that the resources associated with one slice are not used by another [13].

The deployment of these nodes should be carefully planned so that the low-latency and high-reliability requirements can be met simultaneously, since the replacement of a failed MEC by another in a different location may lead to a violation of latency requirements [14], [15]. To address this challenge, the present paper investigates the problem of where to deploy MECs and slices, while respecting Service Level Agreements (SLA), but at minimum cost. Hereinafter, we will use the terms MECs and MEC nodes interchangeably.

Ultra reliability can be supported by various protection schemes based on redundancy, such as $1 : 1$ and $1 : N$. Each scheme imposes different constraints on the location of the MECs but also empowers the network with a specific capacity for the support of reliability. In a previous paper [16], we have investigated the benefits and limitations of these two schemes for furnishing 5G services. As expected, the $1 : 1$ scheme can provide higher levels of reliability, but at a much greater cost. On the other hand, the $1 : N$ scheme provides more cost-effective deployment. Capitalizing on previous findings, the present paper investigates another alternative: a $1 : N : K$ protection scheme to enhance the capability of the $1 : N$ scheme.

In the $1 : N$ scheme, a backup (i.e., secondary) slice is shared among N demand points. Only one of the N demand points can occupy the slice at a given time, and, in the case of a MEC failure, service can be denied if the shared secondary MEC is already occupied by one of the other $N - 1$ demand points. In the $1 : N : K$ protection scheme, however, there are $K - 2$ other redundant slices that can serve the demand point, if the designated shared secondary slice is not available, thus increasing the level of protection. These

Manuscript received 11 August 2021; revised 23 December 2021 and 15 April 2022; accepted 20 June 2022. Date of publication 4 July 2022; date of current version 12 October 2022. The associate editor coordinating the review of this article and approving it for publication was M. Tornatore. *(Corresponding author: Nelson L. S. 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 CP 379C, Cape Verde (e-mail: hernani@lrc.ic.unicamp.br).

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

Digital Object Identifier 10.1109/TNSM.2022.3188159

$K - 2$ additional redundant slices are not reserved for a specific number of demand points, although only a single demand point can occupy the slice at any given time. Moreover, neither of two redundant slices of a demand point can be at the same MEC node nor can the primary slice be at the same MEC with another redundant slice of the same demand point.

This paper postulates and solves the MEC and slice location problem, to determine the location of MECs and slices for the provisioning of 5G services protected by a $1 : N : K$ protection scheme. This location problem is a generalization of the capacitated reliable facility location with a failure probability problem (CRFLP) [17], [18]. The model assumes that MEC nodes are facilities with limited capacities. In this paper, the term slice refers to a subset of the computational and communication resources of a MEC bundled in a virtual machine (VM). Slices can be shared by various demand points, depending on the protection scheme adopted, but only a single demand point can use the slice at a given time to execute an SFC (demand). Different network functions, such as Access and Mobility Function (AMF), Session Management Function (SMF), Policy Control Function (PCF), Application Function (AF), Authentication Server Function (AUSF), User Plane Function (UPF), and User Data Management (UDM)) compose these SFCs. Application latency and reliability requirements must be met, even in the event of failures, therefore, MEC nodes need to be positioned so that 5G QoS requirements will be satisfied. The contributions of this paper are the following:

- A cost-effective protection scheme for 5G services;
- A comparison of the traditional $1 : N$ and $1 : N : K$ protection schemes for furnishing ultra-reliable services in 5G networks;

The present paper differs from previous papers by the same authors in the analysis of the adequacy of the $1 : N : K$ scheme for the provisioning of 5G services, as this has not previously been evaluated. In [16], the $1 : 1$ scheme was compared to the $1 : N$ scheme, with the advantages of the $1 : N$ scheme motivating the present assessment of the cost-effectiveness of the $1 : N : K$ scheme. The results of the present paper show the advantage of the greater reliability of the $1 : N : K$ scheme at an affordable cost.

The rest of this paper is organized as follows. Section II discusses related work. Section III introduces the statement of the problem. Section IV formulates the MEC location problem for the $1 : N : K$ protection scheme. Section V discusses the numerical results for various network scenarios. Section VI concludes the paper.

II. Related Work

This section reviews the state-of-the-art in relation to protection schemes for the MEC location problem and reliability solutions for this problem.

MEC-enabled 5G network planning: Various articles [1], [19], [20], [21], [22], [23] have investigated the placement of VNFs in MECs for the realization of 5G use cases. The work in [24] investigates, in an extensive and detailed way, VNF placement strategies, emphasizing potential issues that may disrupt this placement. The paper classifies the existing placement solutions based on the type (online or offline) and reliability-awareness, and discusses metrics and objectives.

In [19], the authors have tackled the problem of placing control plane functions in a federated cloud architecture. The control plane function SGW-C runs on a virtual machine or container instantiated over a federated cloud. A formulation based on Game Theory seeks a fair balance among cost reduction, load on the SGW-C, and flow installation latency.

In [20], the authors have addressed the problem of placing virtual core network functions. Their approach combines the optimization of the virtual network topology with virtual network embedding optimization, considering requirements of latency for control plane related service chains. The proposed model is based on the embedding of core network service chains without the need for a pre-defined virtual network topology.

In [21], the authors have proposed a management architecture for 5G core networks based on Network Function Virtualization (NFV) and Software Defined Networking (SDN). The authors have formulated the workload allocation problem as a cost minimization problem, considering the cost of the bandwidth in backhaul networks, the energy consumption of mobile clouds and edge, and the revenue associated with delay in the backhaul. The results show that the proposed framework and algorithm reduce the network operationing cost.

In [22], the placement of VNFs on a federated cloud has been proposed. Three solutions were proposed, emphasizing different aspects of a multi-objective problem. The first solution aimed at serving User Equipment (UE) with high mobility and avoids S-GW relocation; the second emphasized serving UEs demanding high Quality of Experience, and the third sought a fair trade-off between these two objectives. Results derived via simulation demonstrated the efficacy of these solutions in achieving the effective placement of VNFs.

In [23], a study of application-driven provisioning of SFC over heterogeneous NFV platforms was undertaken to minimize the total cost of SFCs deployment under the constraint of supporting the QoS requirements of all the SFCs. A solution based on layered auxiliary graphs and an integer linear programming (ILP) model was formulated, that provided near-optimal solutions.

In [25], the authors have proposed a data-driven multi-objective optimization framework for 5G network planning. The work in [26] investigated the resource allocation problem for 5G delay sensitive use cases considering the minimization of total power and bandwidth consumption. In [9], a MEC-enabled 5G Internet of Things (IoT) architecture was introduced to analyze the VNF lifecycle challenges related to a latency-based embedding mechanism for the determination of where the VNFs should be instantiated, scaled, migrated, and finished on the basis of traffic patterns. However, this work considered neither MEC failure nor reliability requirements. In [27], an MEC-based cell selection was proposed for 5G networks which would enable 5G UE to select the cell where the throughput would be maximized.

MEC placement: MEC placement promises to push the computation and communication resources from cloud to network edge. However, due to heterogeneity and variability in MEC nodes, a guarantee of high availability is a challenge for the InP. The authors of [28] studied the problem of a redundant placement policy for the deployment of microservice-based applications at the distributed edge, modeling the distributed redundant placement as a stochastic discrete optimization problem and proposed a Genetic Algorithm-based server selection algorithms designed to optimize response time. However, the proposed problem does not consider the 5G requirements of services.

The authors of [29] addressed the problem of provisioning the data plane for MEC-based cellular networks to improve resiliency, reduce latency, and limit mobility-based service disruptions. However, their work did not address protection schemes, nor was the problem formulated as a multi-objective optimization problem.

The authors in [30] investigated the placement of virtual machine replicas to minimize the average response time of applications running on edges with numerous MEC nodes and considered different demands and heterogeneous MEC capacity constraints. In [31], the authors have proposed a combinatorial optimization solution based on multiple k-redundancy for VM placement considering potential server failures. The proposal estimated the minimum number of VMs necessary to protect services, even with k server failures. However, latency was not considered in the solution.

Reliability solutions: In [32], the authors have considered a workload assignment problem on the basis of latency and reliability requirements to decide to which MEC a workload should be assigned. The authors in [33] investigated the tradeoff between reliability and cost for resource provisioning in fog-aided IoT networks. In [34], the authors investigated the resource provisioning problem of edge nodes subject to potential failures. In [35], the authors studied reliability-aware joint VNF chain placement and flow routing optimization in a cloud data center network using an ILP formulation with reliability constraints. The trade-off between reliability, bandwidth, and consumption of computing resources of service chains was also explored.

In [36], the authors investigated the problem of VNF replica placement for cloud computing to improve QoS provisioning and service reliability. In [37], the authors have presented a scheme for resilient NFV systems with high-performance flow migration schemes. In [38], the authors have proposed a heterogeneous backup deployment scheme to deal with reliability issues.

The authors of [39] have proposed a non-convex mixed-integer nonlinear model to maximize the number of tasks offloaded to a UAV-mounted cloudlet, subject to application latency and reliability requirements.

The authors in [40] have proposed an integer linear programming model to furnish primary and secondary VNF instances for reliable service provisioning when users request different network services with different reliability requirements.

In [41], a heuristic based on Queueing Theory was proposed to furnish a backup model to support shared protection for minimizing the unavailability of VNFs upon request.

However, none of these papers [32], [33], [34], [35], [36], [37], [38], [41] proposes a $1 : N$ protection scheme for the MEC location problem, nor do they investigate the impact of the level of redundancy on resource consumption, which is evaluated in the present paper.

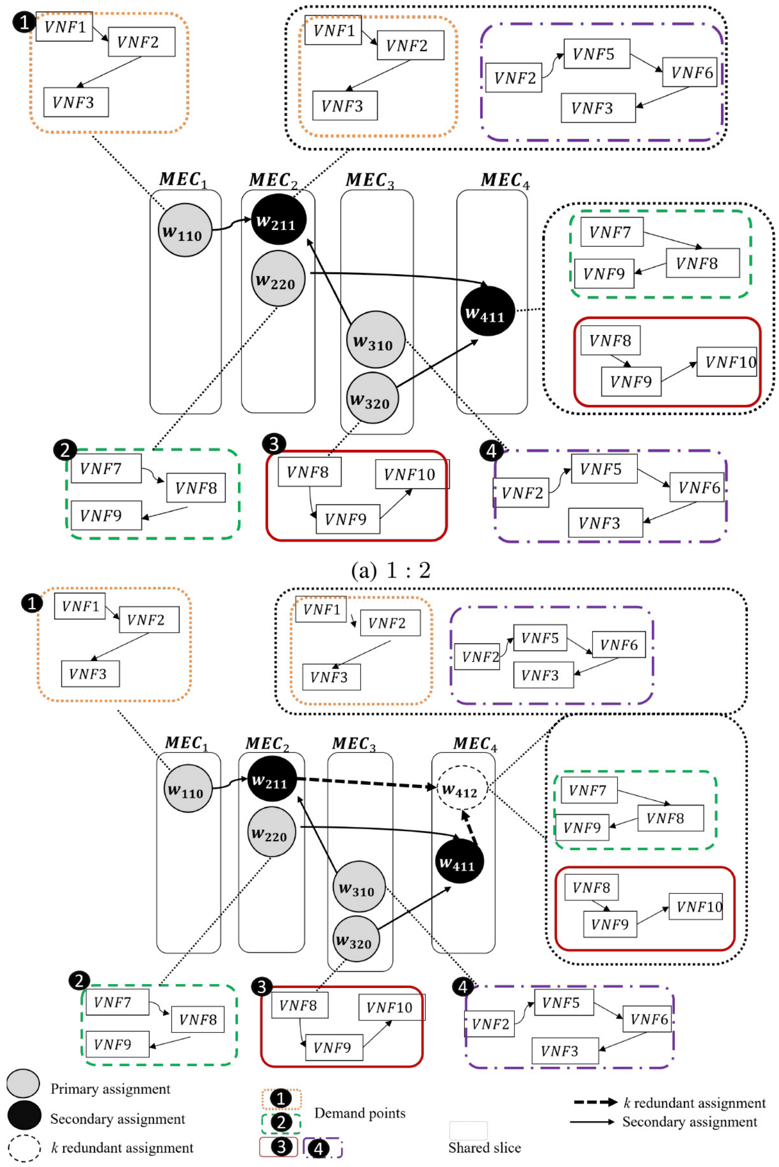
III. Statement of the Problem

Given a set of potential locations, the InP needs to decide to which of them MECs and slices should be deployed to minimize costs, while still satisfying the QoS requirements of 5G services, since MECs have finite computational, storage, and connectivity resources. A subset of computing and network resources of a MEC must be allocated to each isolated slice, with the number of slices hosted by a MEC depending on the resource demanded by the hosted slices. The entire SFC is hosted in a single slice [42] so that the flows between the VNFs do not go through the network links when chaining the VNFs of an SFC, reducing bandwidth consumption. On the other hand, several copies of a VNF type must be placed across the network.

To minimize costs, determined by the number of MECs, slices should be shared by the maximum allowed number of demand points subject to QoS constraints, and these slices should be packed in the minimum number of MECs. This can be achieved by minimizing both the number of slices and the number of MECs employed. The minimization of the number of slices aims at assigning the highest possible number of demand points to a slice. For instance, in a $1 : N$ protection scheme, we aim at having a slice shared by N demand points. Besides that, for a certain number of slices required, the target is to pack them into the minimum possible number of MECs, avoiding the spread of these slices into a number of MECs higher than the minimum possible one. In line with that, it is necessary to minimize both the number of slices and the number of MECs, although these two objective functions are non-conflicting. Moreover, to solve the bi-objective formulation, we employed the lexicographic method, minimizing first the number of slices and then the number of MECs, as justified above.

End-users (called demand points in the Facility Location type of problem) request the execution of SFCs in their exclusive primary slices. A demand point, for example, may solicit the execution of an SFC for Broadcast/Multicast of live streaming, which would be a chain of the vBM-SC, vMBMS-GW, vMME, and vMCE VNFs.

The InP must also choose a protection scheme to support the 5G reliability requirements. Each demand point is associated with a primary (normal operation) slice in a MEC as well as other redundant slices/MECs. In the $1 : N$ protection scheme, for example, N demand points share the same (secondary) slice for protecting their primary slice in the case of a MEC failure. In this case, if the secondary slice is already occupied by one of the other $N - 1$ demand points, the execution of



(a) 1 : 2

(b) 1 : 2 : 3

Fig. 1. Illustration of the 1 : 2 and 1 : 2 : 3 protection schemes.

the SFC of the failed MEC is aborted. No two demand points can simultaneously execute their SFCs in the same slice.

In the $1 : N : K$, however, the SFC executing in a primary slice of a failing MEC can be allocated to another $K - 1$ slices. In this scheme, a demand point is protected by other $K - 2$ slices in addition to the secondary one. In fact, the $1 : N : K$ protection scheme with $K = 2$ is equivalent to the $1 : N$ protection scheme. These $K - 1$ backup slices are ordered for each demand point so that if the k-th node is not available, the SFC will be transferred to the $k + 1$-th slice. Since no execution context is transferred to the new hosting slice, the SFC must again be executed. There is no limit to the number of demand points associated with any one of the $K - 2$ backup slices beyond the secondary one. However, no two slices associated with a demand point can be located at the same MEC. All the backup nodes of a demand point should be located so that the latency requirements are still guaranteed in the case of unavailability of another backup node. If none of the $K - 1$ backup nodes have resources available to host the slices of the failing MEC, the service provided to the demand point is considered lost. The number of backup MECs that should be implemented is a parameter to be evaluated by the InP to guarantee both a target reliability level and cost minimization.

Figure 1 illustrates both the 1 : 2 and the 1 : 2 : 3 protection schemes. In Figure 1, the index number in the VNF representation gives the type of VNF. For instance, the SFC labeled with the number 1 is composed of the execution of *VNF1*, *VNF2*, *VNF3*. Demand points 1, 2, 3 and 4 are assigned exclusively to the primary slices w_{110}, w_{220}, w_{320} and w_{310}, respectively. Demand points 1 and 4 share the secondary slice w_{211}, i.e., they are eligible to use exclusively w_{211} in the case of the

TABLE I
NOTATIONS USED IN THE FORMULATION OF MEC LOCATION PROBLEM

Symbol	Description
Parameters	
U	Set of MECs
V	Set of demand points
E	Set of links
N	Number of demand points that share a secondary slice
K	Protection levels
B_u	Total MEC bandwidth capacity in Mbps
b_v	Bandwidth consumed by the demand point v in Mbps
Ψ_u	MEC CPU capacity in MIPS
Φ_u	MEC RAM capacity in GB
τ_v	Processing demands of the demand point v in MIPS
σ_v	Memory demand of the demand point v in GB
L_v	Bound of the latency requirement of the demand point v in ms
r_u	Failure probability of the MEC $u \in U$
R_v	Required reliability level of the demand point v
Variables	
C_u	Maximum possible number of slices in the MEC u
ψ_{ui}	CPU capacity allocated to the slice i in the MEC u in MIPS
ϕ_{ui}	RAM space allocated to the slice i of the MEC u in GB
l_{uvi}	Latency experienced by demand v when using the slice i of MEC u in ms
y_u	The decision variable indicates whether a MEC u is deployed
w_{uik}	The decision variable indicating whether a MEC u hosts slice i employed as the k-th assignment
x_{uvik}	The decision variable means whether a slice i in the MEC u is employed as the k-th assignment for the demand point v.
Z_{uvik}	Auxiliary continuous decision variable.
ψ^*_{ui}	Auxiliary continuous decision variable.
ψ^U_{ui}	Upper bound of the CPU capacity allocated to the slice i in the MEC u.
ψ^L_{ui}	Lower bound of CPU capacity allocated to the slice i in the MEC u.

failure of their primary slice but they cannot execute their SFC simultaneously in that slice. Moreover, demand points 2 and 3 also share the secondary slice w_{411}. In the 1 : 2 : 3 scheme the w_{412} slice is shared by all the demand points as a redundant slice of the secondary one.

IV. PROBLEM FORMULATION

This section presents a bi-objective formulation for the posed problem. Table I provides the notations used in the paper. A provider can deploy MECs with Ψ_u CPU capacity and Φ_u RAM memory space on the set U of potential locations. The resources of these MECs can be sliced with demand points in the set V associated with specific slices. These slices are allocated for the execution of the SFCs requested by the demand points, as well as the provisioning of protection for the execution of SFCs. The entire SFC is hosted in a single slice. The set E, $E \subseteq U \times V$ represents the links connecting the demand points to potential locations of MECs, and the bipartite graph $G = (U \cup V, E)$ represents the association between demand points and potential MEC locations.

A MEC node can host a maximum number of slices, defined by the ratio between Ψ_u and the smallest capacity required by demand points, $\min_v\{\tau_v\}$. The number of hosted slices is also limited by the ratio between Φ_u and the minimum RAM requirement of the demand points, $\min_v\{\sigma_v\}$. Thus: $C_u = \min(\lfloor \frac{\Psi_u}{\min_v \tau_v}, \frac{\Phi_u}{\min_v \sigma_v} \rfloor)$.

MECs can fail as a consequence of diverse events, such as power outages, and shutdowns. A reliability value, r_u, defines the probability of MEC u and its hosted slices being available.

The following variables will be used to define location decisions:

- $y_u \in \{0, 1\}$ - with the value 1 designating the MEC u as deployed.
- $w_{uik} \in \{0, 1\}$- with the value 1 indicating that the MEC u hosts slice i employed as the k-th assignment ($k = 0$ means primary, $k = 1$ secondary, and $k > 1$ the $K - 2$ additional protection provided by the $1 : N : K$ model beyond the secondary protection).
- $x_{uvik} \in \{0, 1\}$ - with the value 1 meaning that slice i in the MEC u is employed as the k-th assignment for a demand point v.

The MEC location problem can be extended by exploring the employment of the protection scheme $1 : N$ enhanced with $K - 2$ additional redundant MECs to mitigate the impact of failures of a MEC and its hosted slices. The primary slices are allocated to a single demand point, while the secondary slices are shared by $N < |V|$ demand points; $K - 2$ slices are used as redundant to the secondary slice. The $K - 2$ slices redundant to the secondary slice can be shared by any number of demand points. In the event of a failure of the primary slice or hosting MEC, the relevant demand point is reassigned to a secondary slice hosted in a different MEC. Even if the secondary MEC fails, the demand point is still protected by additional $K - 2$ redundant MECs.

The goal is to find the optimal locations for a minimal number of MECs, i.e., to find the minimum number of MECs, which can serve all the demand points and position them appropriately to achieve the goal. Moreover, the number of slices must be minimized. Although the cost of provisioning a slice is not significant when compared to the cost of acquiring a MEC, backup slices should host the maximum possible number of demand points, so that as few additional slices as possible will be demanded, avoiding potential increase in the number of required MECs. These objectives are expressed by (1), (2).

$$Min \sum_{u \in U} \sum_{i=1}^{C_u} \sum_{k=0}^{K-1} w_{uik} \tag{1}$$

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

The constraints of the problem are the following:

$$\sum_{v \in V} x_{uvi0} \leq 1 \quad \forall u \in U, 1 \leq i \leq C_u \tag{3}$$

$$\sum_{v \in V} x_{uvi1} \leq N \quad \forall u \in U, 1 \leq i \leq C_u \tag{4}$$

$$\sum_{i=1}^{C_u} \sum_{k=0}^{K-1} x_{uvik} = 1 \quad \forall v \in V, u \in U \tag{5}$$

$$\sum_{u \in U} \sum_{i=1}^{C_u} x_{uvi0} \leq 1 \quad \forall v \in V \tag{6}$$

$$\sum_{v \in V} x_{uvik} \leq |V| \quad \forall u \in U, 1 \leq i \leq C_u, 2 \leq k \leq K-1 \tag{7}$$

$$\sum_{u \in U} \sum_{i=1}^{C_u} \sum_{k=1}^{K-1} x_{uvik} = K \quad \forall v \in V \tag{8}$$

$$\sum_{k=0}^{K-1} x_{uvik} \leq 1 \quad \forall u \in U, v \in V, 1 \leq i \leq C_u \tag{9}$$

$$x_{uvik} \leq w_{uik} \quad \forall u \in U, v \in V, \\ 1 \leq i \leq C_u, 0 \leq k \leq K-1 \tag{10}$$

$$w_{uik} \leq y_u \quad \forall u \in U, 1 \leq i \leq C_u, 0 \leq k \leq K-1 \tag{11}$$

$$\psi_{ui} = \max_{v \in V, 1 \leq k \leq K-1} \{\tau_v x_{uvik}\} \quad \forall u \in U, \\ 1 \leq i \leq C_u \tag{12}$$

$$\phi_{ui} = \max_{v \in V, 1 \leq k \leq K-1} \{\sigma_v x_{uvik}\} \quad \forall u \in U, \\ 1 \leq i \leq C_u \tag{13}$$

$$\psi_{ui} = \tau_v x_{uvi0} \quad \forall u \in U, v \in V, 1 \leq i \leq C_u \tag{14}$$

$$\phi_{ui} = \sigma_v x_{uvi0} \quad \forall u \in U, v \in V, 1 \leq i \leq C_u \tag{15}$$

$$\sum_{i=1}^{C_u} \psi_{ui} \leq \Psi_u \quad \forall u \in U \tag{16}$$

$$\sum_{i=1}^{C_u} \phi_{ui} \leq \Phi_u \quad \forall u \in U \tag{17}$$

$$\sum_{v \in V} \sum_{i=1}^{C_u} \sum_{k=0}^{K-1} b_v x_{uvik} \leq B_u \quad \forall u \in U \tag{18}$$

$$1 - \prod_{u \in U} \left[(1 - r_u)^{x_{uvi0}}\right] \\ \times \prod_{m \in U \setminus \{u\}} \left[\prod_{k=1}^{K-1} (1 - r_m)^{x_{mvik}}\right] \geq R_v \\ \forall v \in V, u \in U, 1 \leq i \leq C_u \tag{19}$$

$$\sum_{u \in U} \sum_{k=0}^{K-1} \sum_{i=1}^{C_u} \frac{\tau_v}{\psi_{ui}} x_{uvik} \leq L_v \quad \forall v \in V \tag{20}$$

$$y_u, x_{uvik}, w_{uik} \in \{0, 1\}, \\ 0 \leq \phi_{ui} \leq \Phi_u, 0 \leq \psi_{ui} \leq \Psi_u \\ \forall u \in U, v \in V, 0 \leq k \leq K-1, 1 \leq i \leq C_u \tag{21}$$

The objective functions defined by Equations (1) and (2) aim at minimizing the total number of slices and the total number of deployed (active) MECs, respectively.

Constraint (3) indicates that a primary slice is allocated to exactly one demand point. Constraint (4) indicates that a secondary slice is shared by $N < |V|$ demand points. Constraint (5) indicates that the primary slice of a demand point $v \in V$ should not be hosted in the same MEC hosting any other backup slice for this demand point. Constraint (6) indicates that a given demand point $v \in V$ is assigned to exactly one primary slice. Constraint (7) indicates that the MECs used as $K - 2$ redundant for a secondary slice can be shared by all the demand points. Constraint (8) establishes the fact that a given demand point must be protected by $K - 1$ redundant slices.

Constraint (9) indicates that a slice cannot be assigned in more than one level. Constraint (10) indicates that an unassigned slice cannot be allocated to a demand point. Constraint (11) indicates that a slice cannot be hosted in an inactive MEC.

Constraints (12) and (13) indicate, respectively, that the capacity in terms of the CPU and RAM of a *k*-th slice is the maximum capacity value of the N demand points that share that slice. The right-hand sides of the expressions (12) and (13) represent the maximum value of the capacity of all demand points that share that slice. Constraints (14) and (15) state, respectively, that the capacity in terms of CPU and RAM of a primary slice should be greater than or equal to the capacity of the demand point assigned to it. Constraints (16) and (17) ensure, respectively, that the sum of the capacities in terms of CPU and RAM of the slices hosted in a MEC $u \in U$ cannot exceed the capacity of the hosting MEC. Constraint (18) limits the total traffic a MEC *u* can support.

The reliability constraint (19) assures that the reliability level achieved be equal to or greater than R_v, the expected reliability level of a demand point $v \in V$. The left-hand side of the expression (19) is the probability of either the primary slice or a redundant slice hosted on a different MEC being available.

Constraint (20) indicates that the total latency should be less than or equal to 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 time a demand point incurs for processing the *i*-th assigned slice hosted on MEC *u* (l_{uvi}), which is the ratio between the CPU demand τ_v and the CPU capacity allocated to the slice in use, $l_{uvi} = \frac{\tau_v}{\psi_{ui}}$. We assume that when a failure occurs, the execution of the SFC is reinitiated in the MEC to which it has been transferred without any transfer of context. The left-hand side of Constraint (20) considers all MEC to have failed except for the last one (*K* − 1) protecting the demand point. Therefore, it is necessary to ensure that the latency is even lower or at most equal to that which is required in this worst-case scenario of multiple (*K* − 1) failures.

Constraint (21) defines binary decision variables.

The formulation for the 1 : *N* protection scheme can be achieved by making *K* = 2. The formulation in the present paper differs from that in [16] in that response time as an objective (multi-objective) is not minimized in the present paper. Minimizing the response time to values lower than that required would considerably increase the number of demanded MECs, to support 5G QoS requirements. Thus, the formulation introduced in this paper provides the 5G requirements without incurring the unnecessary expenditures encountered in the formulation in [16].

A. Linearization of Problem Formulation

In the previous formulation, Constrains (19) and (20) are non-linear. However, they should be linearized to obtain a completely linear formulation of the problem. Constraint (19) can be linearized by using an exponential function, which yields the following linear constraints.

$$\sum_{u \in U} x_{uvi0} \ln(1 - r_u) + \sum_{m \in U \setminus \{u\}} \sum_{k=1}^{K-1} x_{mvik} \ln(1 - r_m) \\ \leq \ln(1 - R_v) \quad \forall v \in V, 1 \leq i \leq C_u \tag{22}$$

Since the Constraint (20) contains the ratio of a binary decision variable x_{uvik} with a continuous decision variable ψ_{ui}, it is non-linear, but the expression $\frac{1}{\psi_{ui}}$, $\psi_{ui} \in R^{+}$ can be modeled by introducing a continuous decision variable ψ^{*}_{ui} and adding the equation $\psi_{ui}\psi^{*}_{ui} = 1$. Considering this new equation, $\psi^{*}_{ui} = \frac{1}{\psi_{ui}}$, and the bounds $\psi_{ui} \in [\psi^{L}_{ui}, \psi^{U}_{ui}]$ and $\psi^{*}_{ui} \in [\frac{1}{\psi^{U}_{ui}}, \frac{1}{\psi^{L}_{ui}}]$, Constraint (20) can be rewritten as:

$$\sum_{u \in U} \sum_{k=0}^{K-1} \sum_{i=1}^{C_u} \tau_v \psi^{*}_{ui} x_{uvik} \le L_v \quad \forall v \in V$$

$$\psi_{ui} \in [\psi^{L}_{ui}, \psi^{U}_{ui}] \quad \psi^{*}_{ui} \in \left[\frac{1}{\psi^{U}_{ui}}, \frac{1}{\psi^{L}_{ui}}\right] \tag{23}$$

The expression (23) is still non-linear, however, since it includes the product $\psi^{*}_{ui} x_{uvik}$ of a binary decision variable $x_{uvik} \in \{0, 1\}$ with a continuous decision variable $\psi^{*}_{ui} \in [\frac{1}{\psi^{U}_{ui}}, \frac{1}{\psi^{L}_{ui}}]$. To linearize the expression (23), we introduce an auxiliary continuous decision variable $Z_{uvik} \in \mathbb{R}^{+}$, such that $Z_{uvik} = \psi^{*}_{ui} x_{uvik}$, yielding the following new expression:

$$\sum_{u \in U} \sum_{k=0}^{K-1} \sum_{i=1}^{C_u} \tau_v z_{uvik} \le L_v \quad \forall v \in V$$

$$z_{uvik} \le \frac{1}{\psi^{L}_{ui}} x_{uvik}$$

$$z_{uvik} \ge \frac{1}{\psi^{U}_{ui}} x_{uvik}$$

$$z_{uvik} \le \psi^{*}_{ui} + \frac{1}{\psi^{U}_{ui}} (1 - x_{uvik})$$

$$z_{uvik} \ge \psi^{*}_{ui} - \frac{1}{\psi^{L}_{ui}} (1 - x_{uvik})$$

$$\psi_{ui} \in \left[\psi^{L}_{ui}, \psi^{U}_{ui}\right]$$

$$\psi^{*}_{ui} \in \left[\frac{1}{\psi^{U}_{ui}}, \frac{1}{\psi^{L}_{ui}}\right] \tag{24}$$

By replacing Constraints (19) and (20) with Constraints (22) and (24), the previous non-linear formulation has been transformed into a linear one, which can now be solved with an optimization solver.

B. Bi-Objective Metaheuristic

The MEC location problem with a protection scheme is a constrained bi-objective optimization NP-hard problem. The problem of locating a reliable MEC can be formulated as a generalization of the classical facility location problem [17], known as the capacitated reliable facility location problem with failure probability (CRFLP) [18]. This problem therefore is NP-hard, since it is a generalization of the NP-hard capacitated facility location with failure probability problem. The metaheuristic non-dominated sorting genetic algorithm (NSGA)-II [43] was then employed to solve the proposed formulation. The advantage of using (NSGA)-II for this problem has been described in previous work [44].

This algorithm uses ranking and crowding criteria to try to find a set of solutions that are not dominated (Pareto front) by any other solution. A dominant solution is one that is considered better than all others. Solution *i* dominates solution *j*, if solution *i* is better than or equal to solution *j* on the basis of the adopted number of activated MECs, reliability, and the number of hosted slices. Three different ranking techniques have been used to find a non-dominated solution. The first ranking uses the objective function values, the second considers all constraint violations, and the third combines the two criteria. After ranking, the solutions at the top of the ranking are chosen.

The NSGA-II algorithm tailored to solve the MEC Location Problem starts from an initial network design containing a set of deployed (activated) MECs and hosted slices which are assigned to demand points. A tournament selection is then realized by selecting parent networks and creating child networks using crossover and mutation. The crossover operation involves the random selection of two individuals from the population and production of offspring inheriting as much useful information as possible from the two individuals. The mutation operation is then applied to each gene using a mutation probability value. The mutated genes generate a new value to produce a new population, emulating the creation of slices for each active MEC, thus resulting in an enhanced solution for the MEC location problem. If the individual created is not valid, i.e., the algorithm cannot derive a sufficient number of slices for hosting on an active MEC to fulfill the reliability requirements, the chromosomes are discarded and a new individual using the crossover and mutation operations is created.

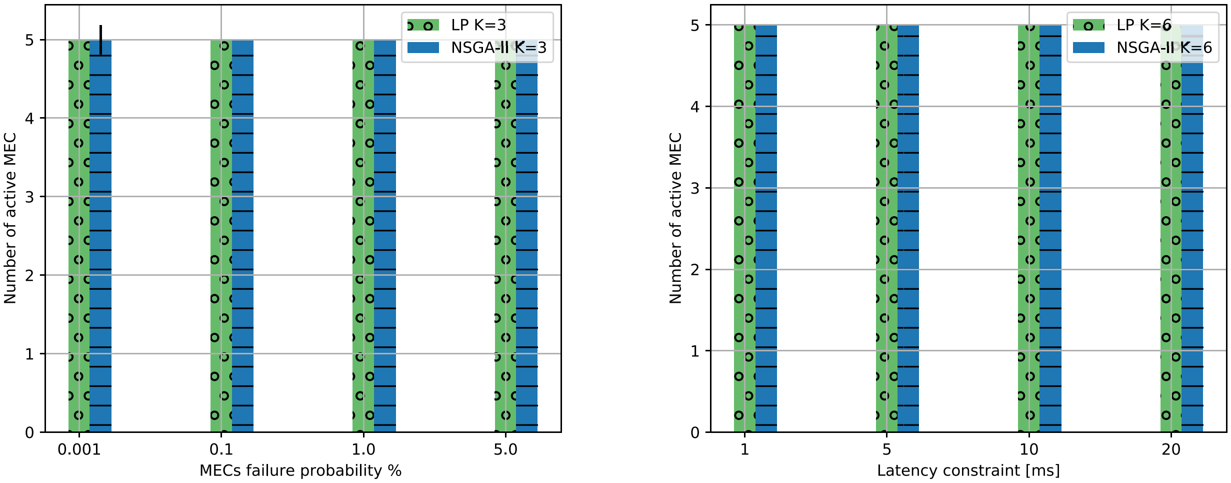
These next-generation networks are selected after applying non-dominated and crowding distance sorting methods. In non-dominated sorting, networks are sorted based on their ability to support latency and reliability requirements as well as the number of activated MECs and the hosted slices. A population of networks with the best configuration is then selected on the basis of the achievement of objectives and respect for the constraints (fitness value). 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.

V. Evaluation of the $1 : N : K$ Protection Scheme

This section presents the results obtained for different scenarios for the comparison of the schemes $1 : N$ and $1 : N : K$ in the evaluation of the solution proposed for the MEC Location Problem. The source code for the proposed optimization problem is available at https://github.com/hdchantre/jMetalPy [45].

A. Scenario

Following the work in [46], the infrastructure scenario for the MEC location problem is composed of MEC nodes, distributed in a grid topology over an area of 1000 x 1000 meters. The designed network infrastructure is composed of 20 MEC nodes with 8 virtual CPUs MIPS (4800), and 8GB of RAM. The data rate supported by the MECs is 400 Mbps [47]. The population size for the algorithm was set to 100, the number of generations to 25, and the confidence level to 95%.



(a) Number of MECs demanded for U=5 as a function of MEC failure probability.

(b) Number of MECs demanded for U=5 as a function of latency constraints.

Fig. 2. Number of MEC demanded for different scenarios.

Demand points request the execution of SFCs (carried out in a slice of a MEC node). A single SFC is associated with each demand point, and only one SFC can be hosted in a slice at a certain time. The number of VNFs of an SFC was randomly determined from a Uniform distribution in the interval [1, 10]. Each VNF requires processing, storage, and bandwidth demands, which are uniformly distributed in the range [0, 4] vCPUs, [0, 6] RAM, and [100, 300] Mbps, respectively. The demand values of an SFC are the maximum value of the demands of its VNFs. An SFC also has reliability and latency requirements. Latency requirements are classified as latency-sensitive ($L_v = \{1; 5\}ms$) or latency tolerant ($L_v = \{10; 20\}ms$). Four classes of reliability requirements were identified to represent different reliability-aware service demands: high-level (99.999%), middle-level (99.99%) and (99.9%), and low-level (99.0%). The level of redundancy for the hosted slices was also varied: low-level ($K = 2$, $K = 3$), middle level ($K = 4$, $K = 5$), and high-level ($K = 6$, $K = 7$). Results of the $1 : N$ scheme are represented with the level of reliability set to $K = 2$. Moreover, $|U| = 20$, $|V| = 1000$, and $N = 20$.

B. Evaluation of Metaheuristic

This section assesses the precision of the proposed metaheuristic by comparing the results with those given by the bi-objective constrained linearized integer formulation introduced in Section IV-A.

All results in the paper were derived using a computer equipped with JMetal version 5.3 [48], Debian GNU/Linux Squeeze, two Intel Xeon (2.13GHz) with 4 cores each, and 78GB RAM.

The bi-objective constrained formulation is an NP-hard problem, so all the feasible instances of this formulation were solved using the Gurobi Optimizer solver. To solve the bi-objective formulation, we first minimized the number of slices (Equation (1)) and then minimize the number of MECs (Equation (2)). The parameters $|U|, |V|, K, N$ were varied to produce feasible instances. The maximum values of $|U|, |V|, K, N$ were 5, 9, 6 and 5, respectively. The results were derived as a function of the reliability requirement, probability of failure, and latency requirements.

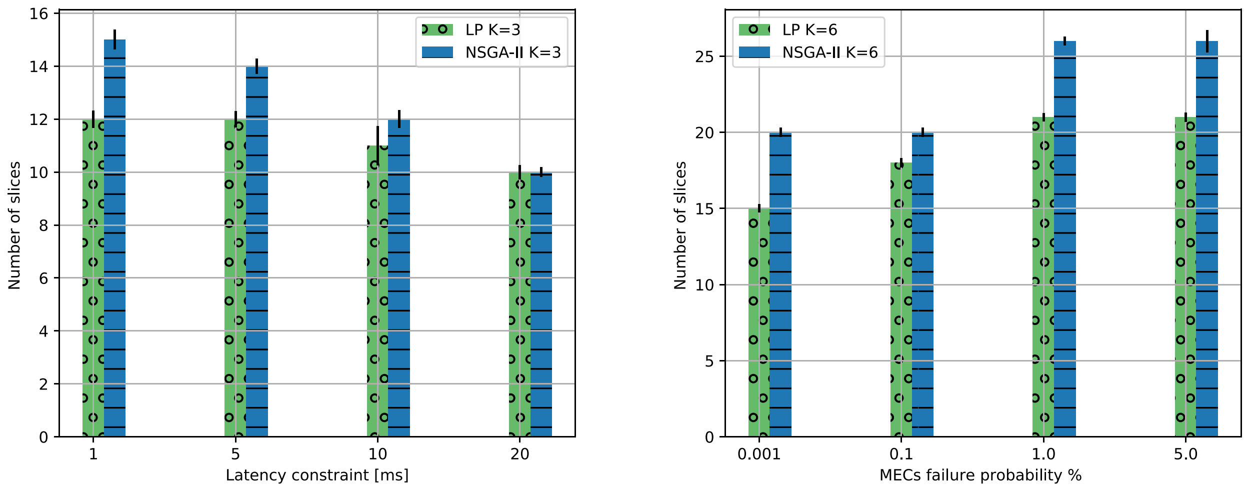
Figures 2a and 2b show the number of MECs demanded for different scenarios. The number of MECs suggested by the metaheuristic was the same as that suggested by the bi-objective constrained linearized integer formulation for all scenarios evaluated.

However, the metaheuristic overestimates the number of demanded slices when compared to the number predicted by the bi-objective constrained linearized integer formulation. The greater the level of redundancy (K), the greater is the overestimation. It was up to 25% greater for all the scenarios evaluated. Figures 3a and 3b show the number of demanded slices for the different scenarios.

The metaheuristic was used to evaluate larger scenarios for which it was impossible to derive a solution by employing bi-objective constrained linearized integer formulation due to the extremely high computational demands required to obtain a solution with existing resources and in a timely manner.

C. Evaluation of the Need of Bi-Objective Formulation

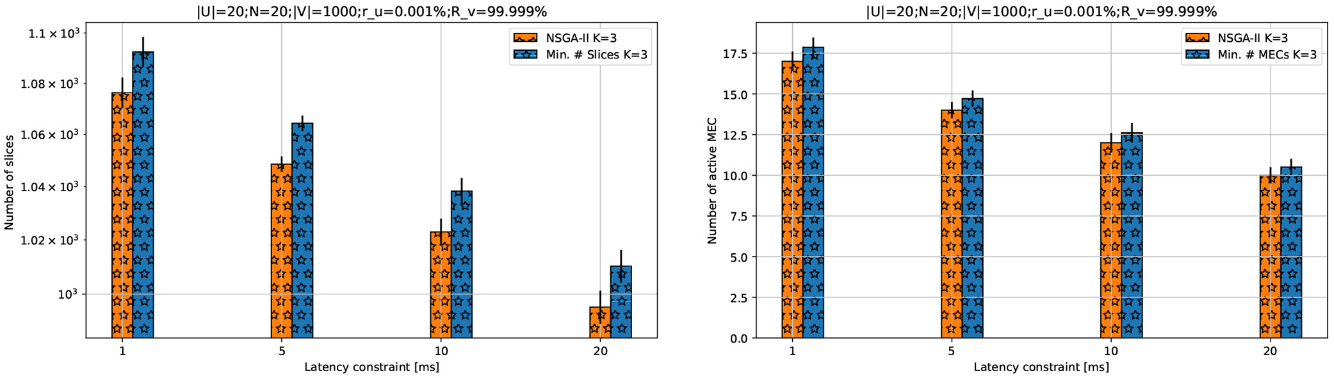
In order to evaluate the necessity of having a bi-objective formulation rather than a mono-objective one, we compared the results given by the proposed bi-objective formulation with those given by two mono-objective ones, each with one of the proposed objective functions. We compared the linear programming formulations for small scenarios and the metaheuristic formulations for larger scenarios. Figures 4a and 4b show the number of slices and MECs demanded by the bi-objective and mono-objective metaheuristic formulation as a function of the latency constraint. These figures show that when the problem is modeled with a single objective, the number of slices and MECs demanded will be greater than those indicated by the bi-objective formulation. Figures 5a and 5b show the results derived using



(a) Number of Slices demanded for U= 5 as a function of latency constraints.

(b) Number of Slices demanded for U=5 as a function of the MEC failure probability.

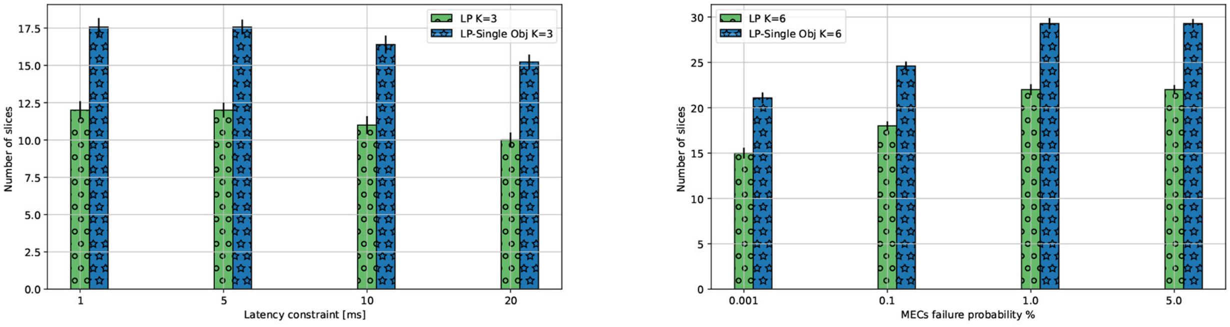
Fig. 3. Number of Slices demanded for different scenarios.



(a) Number of Slices demanded as a function of latency constraints.

(b) Number of MECs demanded as a function of latency constraints.

Fig. 4. Number of Slices and MEC demanded.

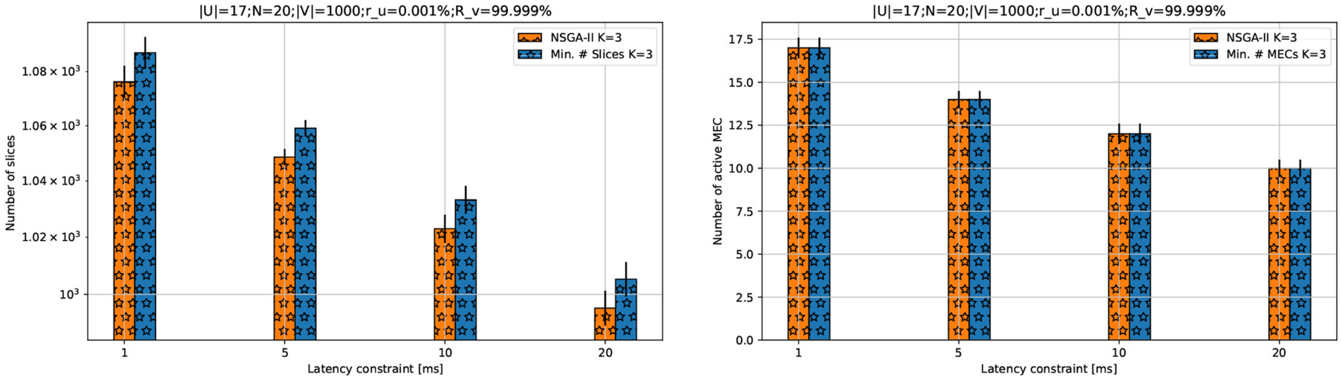


(a) Number of Slices demanded for U= 5 as a function of latency constraints.

(b) Number of Slices demanded for U=5 as a function of the MEC failure probability.

Fig. 5. Number of Slices demanded.

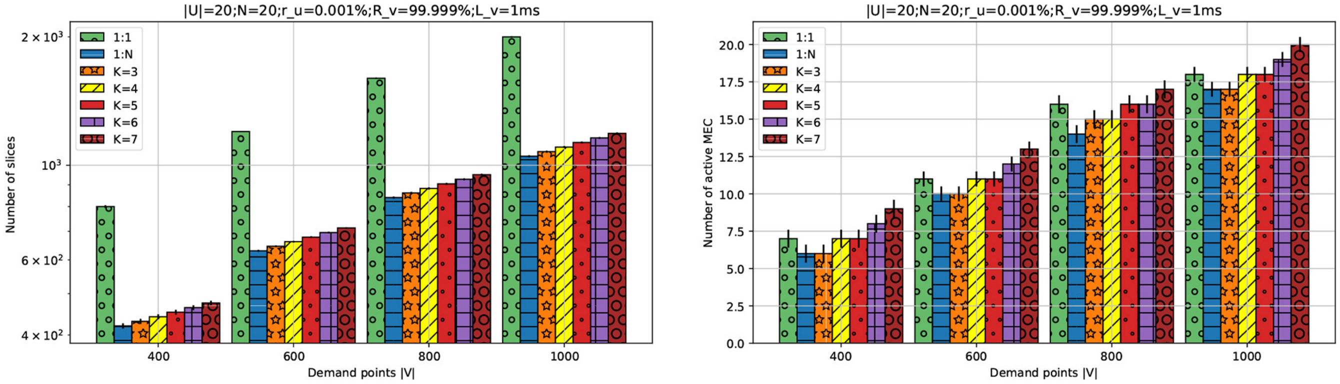
the linear programming formulation for a small scenario. This small scenario demanded a MEC in every potential location, and the mono-objective formulation demanded a greater number of slices. We also varied the size of the scenario to identify the size, which calls for a bi-objective formulation. We found that, when $|U| \leq 17$, the number of MECs demanded is the same as a result of the need to have a MEC in every location, but the mono-objective formulation always demands a higher number of slices (Figures 6a and 6b)



(a) Number of Slices demanded as a function of latency constraints.

(b) Number of MECs demanded as a function of latency constraints.

Fig. 6. Number of Slices and MEC demanded.



(a) Number of Slices required.

(b) Number of MECs required.

Fig. 7. Number of slices and MECs required as a function of demand points.

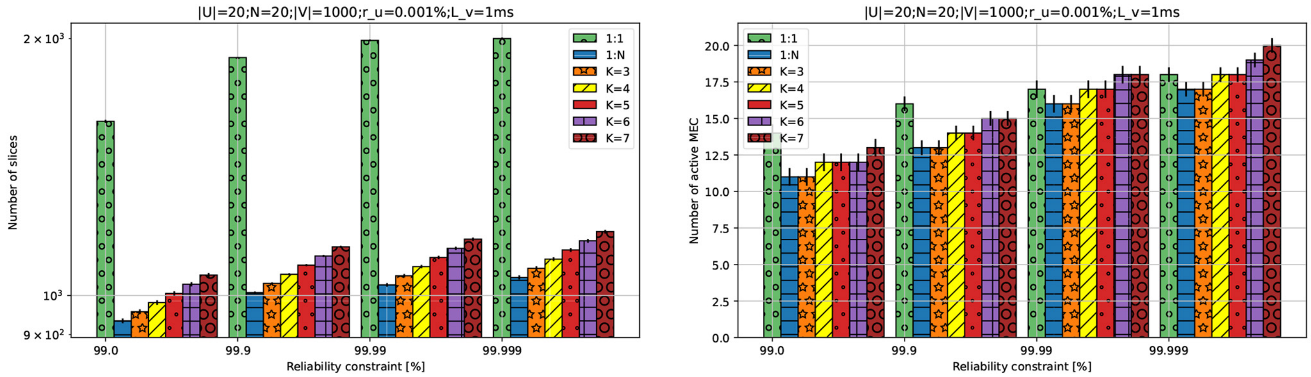
D. Numerical Evaluation

This section shows the design of protected networks with different protection schemes as a function of the number of demand points, reliability requirements, latency requirements, and level of sharing of the secondary backup (N). The figures present the demanded number of slices and MECs.

Figures 7a and 7b show the number of slices and MECs needed as a function of the number of demand points, respectively. Figure 7a shows that the $1:1$ protection scheme demands more slices than do the others protection schemes ($1:N$ and $1:N:K$) because of the exclusive allocation of the secondary backup. The $1:N$ scheme demands the fewest slices. As the level of redundancy K increases, so does the number of required slices. This small increase is the result of the fact that there is no limit to the number of demand points sharing a slice for $K \geq 3$. The compliance with latency requirements is responsible for most of the increase in the number of slices as a function of the level of protection. The number of slices increases almost linearly as a function of the number of demand points $|V|$.

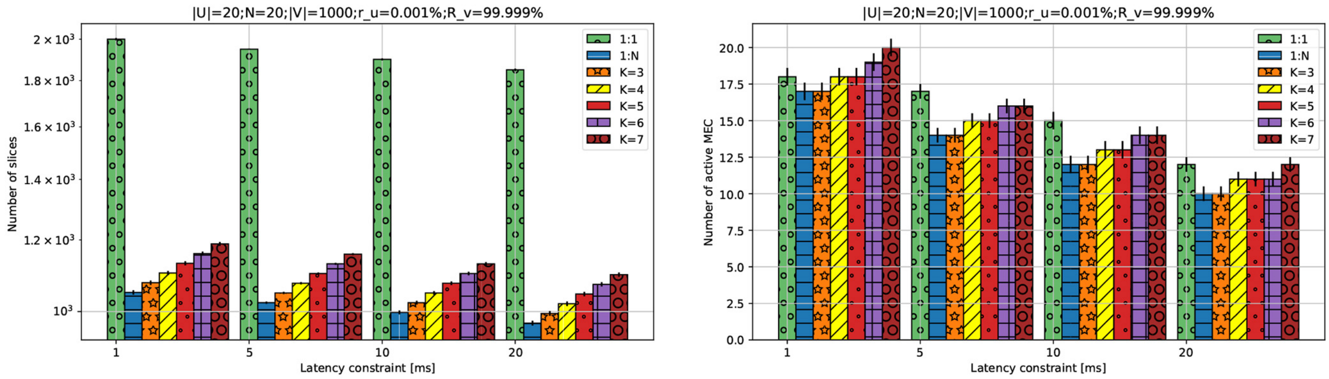
Figure 7b depicts that the required number of MECs does not necessarily follow the same trend of the number of slices required. The high demand for slices in the $1:1$ protection scheme does not imply a much higher demand for MECs than do the other protection schemes. The number of MECs demanded by the $1:1$ scheme is similar to that of $K = 5$ and always higher than that of $K = 3$. The number of MECs required by protection schemes other than $1:1$ is impacted by the requirement that no two protective slices for a demand point can reside in the same MEC, and it increases as a function of the protection level (K). Moreover, the latency requirement also impacts the number of required MECs since additional MECs may be necessary to cope with more stringent latency requirements. In most cases, $1:N$ and $1:N:3$ demand the same number of MECs, and the same is also true for $K = 4$ and $K = 5$. Increasing the protection level does not necessarily increase the number of demanded MECs, since for $K \geq 3$ there is no restriction on the number of demand points that can share a slice.

Figures 8a and 8b show the number of required slices and MECs as a function of reliability requirements, respectively. Results are shown considering a strict service latency requirement of $L_v = 1ms$, a probability of failure of $r_u = 0.001\%$, and $|V| = 1000$. Hereinafter, $|V| = 1000$ for all the figures. Fig. 8a confirms the pattern seen in Fig. 7a. The $1:1$ protection scheme demands more slices than do the other protection schemes. The greater the level of protection, the higher is the number of demanded slices. By increasing the protection



(a) Number of slices demanded as function of reliability requirements. (b) Number of MECs demanded as a function of reliability requirements.

Fig. 8. Number of slices and MECs demanded as a function of reliability requirements.



(a) Number of Slices demanded as function of latency constraints. (b) Number of MECs demanded as a function of latency constraints.

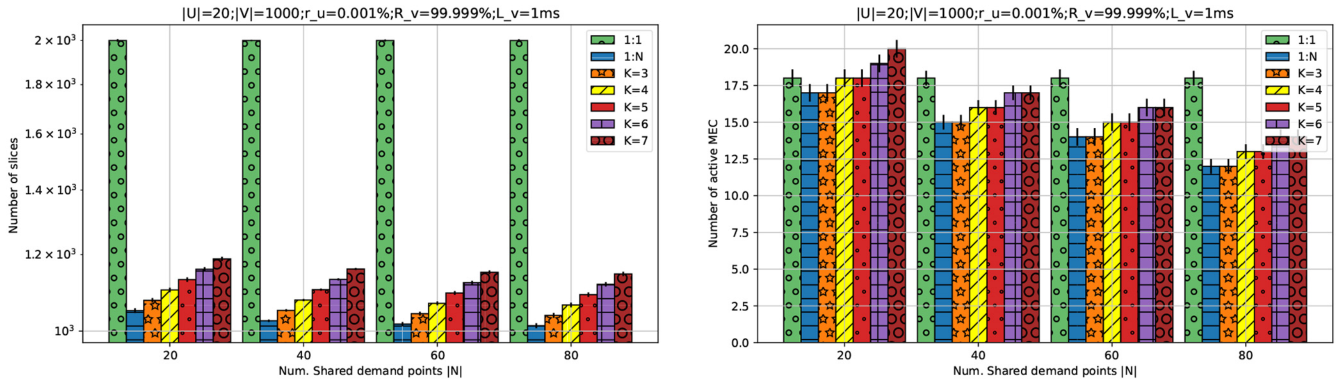
Fig. 9. Number of Slices and MECs demanded as a function of latency constraints.

levels from $K = 3$ to $K = 4$, $K = 5$, $K = 6$, and $K = 7$, the number of slices increases by 2.5%, 5%, 10%, 12.5%, respectively. These results demonstrate that the number of demanded slices increases almost linearly with reliability requirements. The increase in the number of slices as a function of the level of protection (K) is not as great, since there is no limit to the number of demand points per slice. The results show that for the 1 : N : 3 protection scheme, the addition of a '9' to the reliability requirement (from 99.0% to 99.9%, from 99.9% to 99.99%, and from 99.99% to 99.999%) implies an increase of 74, 21, 21 slices, respectively.

Fig. 8b shows the impact of the reliability requirement on the number of demanded MECs. The number of MECs increases almost linearly with an increase in the reliability requirement. For the $1 : N$ and 1 : N : 3 protection schemes, the number of demanded MECs is 11, 13, 16, and 17 for the reliability levels 99.0% 99.9%, 99.99%, 99.999%, respectively. The number of demanded MEC does not necessarily increase with the protection level; the protection schemes $1 : N$ and $1 : N : 3$ demanded the same number of MECs as did those for $K = 4$ and $K = 5$. This evinces that it is possible to offer an additional level of protection without increasing costs. This is dependent on the SFCs resource demands and the MEC resource capacity. Such a possibility also depends on the reliability constraints. The protection levels $K = 6$ and $K = 7$ differ when the constraint is 99.999, although they were the same for the constraints of 99.9 and 99.99.

Figures 9a and 9b consider the impact of latency requirements on the design of edge networks, with both latency-sensitive ($L_v = 1; 5ms$) and latency tolerant ($L_v = 10; 20ms$) requirements being considered. A five-nines reliability requirement, $R_v = 99.999\%$, with a probability of failure of $r_u = 0.001\%$ were considered here. Fig. 9a and 9b show, respectively, that the demand for slices and MECs decreases when less stringent requirements are considered, i.e., the latency values increase. The number of slices decreases by 8% when the latency requirements increase from $L_v = 1ms$ to $L_v = 20ms$, while the number of demanded MECs decreases with less stringent latency requirements. The more strict the latency requirement, the greater is the demand for MECs. This increase in the number of MECs is a function of K, since the protective slices of a demand point must be hosted in different MECs.

Fig. 10a and 10b evaluate the impact of the number of demand points $|N|$ sharing a secondary backup slice on the number of demanded slices and MECs for the $1 : 1$, $1 : N$ and $1 : N : K$ protection schemes. A strict service latency requirement of $L_v = 1ms$ and a probability of failure of $r_u = 0.001\%$ were employed. The results in Fig. 10a and Fig. 10b confirm that as the number of shared demand points $|N|$ increases,



(a) Number of Slices demanded .

(b) Number of MECs demanded.

Fig. 10. Number of Slices and MECs demanded as a function of shared demand points $|N|$.

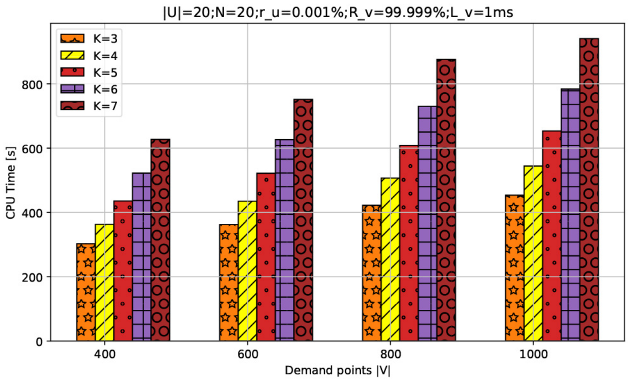


Fig. 11. CPU Time as function of demand points $|V|$.

the number of demanded slices and MECs to be deployed decreases. The number of deployed MEC was the same for both $K = 3$ and $1 : N$ protection schemes, regardless of the value of N.

Figure 11 shows the impact of the number of demand points and the protection level on the CPU time required to execute the proposed heuristics. The increase in computing time is linear in relation to the number of demand points. The CPU time also increases as the protection level K increases, as well as that of the number of constraints.

Providing for an extra level of redundancy should be carefully considered by the InP, as this can lead to additional benefits with little or no additional MEC required. When the requirements for reliability are high, an increase in the level of protection can be crucial for the provisioning of protected services. Overall, the number of MECs demanded by the level of redundancy $K = 3$ and $1 : N$ is the same, and in most cases, this is also true when $K = 4$ and $K = 5$.

VI. Conclusion

This paper has investigated the MEC Location Problem with a $1 : N : K$ protection scheme, which offers additional redundancy in the case of MEC failure beyond the secondary backup MEC shared by N demand points. The provisioning of stringent requirements is a challenge for the InP, but they can be achieved by adequate positioning of the MECs in the InP infrastructure. Latency requirements must be supported, even in the case of a MEC failure and the transfer of the SFC to a redundant slice. Since redundancy beyond the secondary backup node is achieved by MECs shared by any number of demand points, the cost of providing an additional level of protection is not necessarily much higher than that for secondary slice shared by only N demand points. This has been clearly shown for various scenarios where no significant increase in number of demanded MECs was observed with an increase in protection. We have, thus, postulated that the $1 : N : K$ scheme provides various advantages for reliable service provisioning in 5G.

This paper has investigated the design problem for different requirements of reliability and latency, and has shown that relaxing the stringent five-nines reliability requirement can lead to a significant reduction in cost. This is also true for a low-latency requirement, established as 1 ms for initial efforts of 5G standardization. The formulation introduced in this paper can help InPs plan their infrastructure according to their use case demands. Differentiated reliability is one way of dealing with diverse reliability requirements, and the investigation of this approach for the provisioning of 5G services is recommended.

The placement of the VNFs of an SFC is usually guided by the choice of the performance metrics to optimize [49]. Here, we have made the assumption that an SFC is fully hosted by a single MEC so that network delays for the completion of an SCF are nullified and service latency minimized so that strict latency requirements of 5G use cases can be supported. As future work, we plan to consider the case of VNFs of a service chain placed on different MECs. For that, the topology of the network connecting the MECs should be considered in the problem formulation, as well as the capacity of the network links. Moreover, VNFs should be indexed in the decision variables of the problem formulation. Such a study should be able to assess the extent to which the overhead of network delay in the execution of an SFC impacts the provisioning of reliability for 5G services.

Acknowledgment

The authors would like to thank PEC-PG-CAPES/CNPq Brazil for the financial support. We would also like to acknowledge grant #15/24494-8 from the Sao Paulo Research Foundation (FAPESP) and CNPq.

References

[1] J. Yusupov, A. Ksentini, G. Marchetto, and R. Sisto, "Multi-objective function splitting and placement of network slices in 5G mobile networks," in *Proc. IEEE Conf. Stand. Commun. Netw. (CSCN)*, 2018, pp. 1–6.

[2] P. Dinh, M. A. Arfaoui, S. Sharafeddine, C. Assi, and A. Ghrayeb, "A low-complexity framework for joint user pairing and power control for cooperative NOMA in 5G and beyond cellular networks," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6737–6749, Nov. 2020.

[3] 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)*, 2018, pp. 1–7.

[4] M. Savi, M. Tornatore, and G. Verticale, "Impact of processing-resource sharing on the placement of chained virtual network functions," *IEEE Trans. Cloud Comput.*, vol. 9, no. 4, pp. 1479–1492, Oct.–Dec. 2021.

[5] 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 architecture & orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[6] B. Li, W. Lu, and Z. Zhu, "Deep-NFVOrch: Deep reinforcement learning based service framework for adaptive VNF service chaining in IDC-EONs," in *Proc. Opt. Fiber Commun. Conf. Exhibit. (OFC)*, 2019, pp. 1–3.

[7] S. Yu, X. Chen, Z. Zhou, X. Gong, and D. Wu, "When deep reinforcement learning meets federated learning: Intelligent multi-timescale resource management for multi-access edge computing in 5G ultra dense network," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2238–2251, Feb. 2021.

[8] 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.

[9] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P. Mekikis, A. Antonopoulos, and C. Verikoukis, "Online VNF Lifecycle management in an MEC-enabled 5G IoT architecture," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4183–4194, May 2020.

[10] A. Banerjee, N. Sastry, and C. M. Machuca, "Sharing content at the edge of the network using game theoretic centrality," in *Proc. 21st Int. Conf. Transp. Opt. Netw. (ICTON)*, 2019, pp. 1–4.

[11] P. 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.

[12] 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./Oct. 2019.

[13] A. J. Gonzalez *et al.*, "The isolation concept in the 5G network slicing," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, 2020, pp. 12–16.

[14] D. Santos, T. Gomes, and D. Tipper, "Software-defined network design driven by availability requirements," in *Proc. 16th Int. Conf. Design Rel. Commun. Netw.*, 2020, pp. 1–7.

[15] S. Ayoubi, Y. Chen, and C. Assi, "Towards promoting backup-sharing in survivable virtual network design," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 3218–3231, Oct. 2016.

[16] H. D. Chantre and N. L. S. da Fonseca, "The location problem for the provisioning of protected slices in NFV-based MEC infrastructure," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 7, pp. 1505–1514, Jul. 2020.

[17] C. Aikens, "Facility location models for distribution planning," *Eur. J. Oper. Res.*, vol. 22, no. 3, pp. 263–279, 1985.

[18] R. Yu, "The capacitated reliable fixed-charge location problem: Model and algorithm," M.S. thesis, Dept. Ind. Syst. Eng., Lehigh Univ., Bethlehem, PA, USA, 2015.

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

[20] 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.

[21] 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.

[22] 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)*, 2015, pp. 3879–3884.

[23] L. Dong, N. L. S. da Fonseca, and Z. Zhu, "Application-driven provisioning of service function chains over heterogeneous NFV platforms," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 3, pp. 3037–3048, Sep. 2021.

[24] 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.

[25] B. B. Haile, E. Mutafungwa, and J. Hämäläinen, "A data-driven multiobjective optimization framework for hyperdense 5G network planning," *IEEE Access*, vol. 8, pp. 169423–169443, 2020.

[26] M. Tohidi, H. Bakhshi, and S. Parsaeefard, "Flexible function splitting and resource allocation in C-RAN for delay critical applications," *IEEE Access*, vol. 8, pp. 26150–26161, 2020.

[27] S. Natarajan, T. Khandelwal, and M. Mittal, "MEC enabled cell selection for micro-operators based 5G open network deployment," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, 2020, pp. 1–5.

[28] H. Zhao, S. Deng, Z. Liu, J. Yin, and S. Dustdar, "Distributed redundancy scheduling for microservice-based applications at the edge," *IEEE Trans. Services Comput.*, vol. 15, no. 3, pp. 1732–1745, May/Jun. 2022.

[29] R. Ford, A. Sridharan, R. Margolies, R. Jana, and S. Rangan, "Provisioning low latency, resilient mobile edge clouds for 5G," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2017, pp. 169–174.

[30] L. Zhao and J. Liu, "Optimal placement of virtual machines for supporting multiple applications in mobile edge networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6533–6545, Jul. 2018.

[31] F. Machida, M. Kawato, and Y. Maeno, "Redundant virtual machine placement for fault-tolerant consolidated server clusters," in *Proc. IEEE Netw. Oper. Manage. Symp.*, 2010, pp. 32–39.

[32] N. Kherraf, S. Sharafeddine, C. M. Assi, and A. Ghrayeb, "Latency and reliability-aware workload assignment in IoT networks with mobile edge clouds," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 4, pp. 1435–1449, Dec. 2019.

[33] J. Yao and N. Ansari, "Fog resource provisioning in reliability-aware IoT networks," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8262–8269, Oct. 2019.

[34] L. Toka, D. Haja, A. Kőrösi, and B. Sonkoly, "Resource provisioning for highly reliable and ultra-responsive edge applications," in *Proc. IEEE 8th Int. Conf. Cloud Netw. (CloudNet)*, 2019, pp. 1–6.

[35] 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.

[36] 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.

[37] 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.

[38] 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)*, 2018, pp. 1096–1103.

[39] E. E. Haber, H. A. Alameddine, C. Assi, and S. Sharafeddine, "A reliability-aware computation offloading solution via UAV-mounted cloudlets," in *Proc. IEEE 8th Int. Conf. Cloud Netw. (CloudNet)*, 2019, pp. 1–6.

[40] M. Huang, W. Liang, X. Shen, Y. Ma, and H. Kan, "Reliability-aware virtualized network function services provisioning in mobile edge computing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 11, pp. 2699–2713, Nov. 2020.

[41] F. He and E. Oki, "Unavailability-aware shared virtual backup allocation for Middleboxes: A queueing approach," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 2, pp. 2388–2404, Jun. 2021.

[42] D. Li, P. Hong, K. Xue, and J. Pei, "Virtual network function placement and resource optimization in NFV and edge computing enabled networks," *Comput. Netw.*, vol. 152, pp. 12–24, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128618305000

[43] 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.

[44] 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.

[45] "Source Code." 2021. [Online]. Available: https://github.com/hdchantre/jMetalPy

[46] 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.

[47] V. Scoca, A. Aral, I. Brandic, R. D. 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.

[48] A. J. Nebro, J. J. Durillo, and M. Vergne, "Redesigning the jMetal multi-objective optimization framework," in *Proc. Compan. Publ. Annu. Conf. Genet. Evol. Comput.*, 2015, pp. 1093–1100.

[49] D. Bhamare, R. Jain, M. Samaka, and A. Erbad, "A survey on service function chaining," *J. Netw. Comput. Appl.*, vol. 75, pp. 138–155, Nov. 2016.

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, Bridgewater, MA, USA, in 2010. He is currently pursuing the Ph.D. degree in computer science with the State University of Campinas, Brazil. He is an Assistant Graduate Professor with the University of Cape Verde. His research interests include network function virtualization and network-based cloud computing.

Nelson L. S. da Fonseca (Senior Member, IEEE) 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 with the Institute of Computing, State University of Campinas, Campinas, Brazil. He has published 450+ papers and supervised over 70 graduate students. He is the recipient of the 2020 ComSoc Harold Sobol Award for Exemplary Service to Meetings and Conferences, the 2012 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 Conferences of the IEEE Communications Society (ComSoc). He also served as the ComSoc Vice President Technical and Educational Activities, the Vice President Publications, the Vice President Member Relations, and the Director of Conference Development, Latin America Region, and On-Line Services. He is the Past Editor-in-Chief of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.