# Multi-Objective Optimization for Edge Device Placement and Reliable Broadcasting in 5G NFV-Based Small Cell Networks

Hernani D. Chantre and Nelson L. S. da Fonseca

***Abstract*— This paper investigates the problem of locating edge devices in ultra-dense 5G network function virtualization-based small cell networks for the provisioning of reliable broadcasting services. The problem is formulated as a capacitated reliable facility location problem (CRFLP) with failure probability. The solution aims at placing the VNFs of broadcast transmissions optimally on selected edge devices in order to ensure high reliability and minimize the cost of providing broadcast services as well as the probability of loss of service requests. The CRFLP is a non-linear NP-hard problem and it is evaluated by using a multi-objective evolutionary algorithm. Two multi-objective metaheuristics are compared: the multi-objective particle swarm optimization and the nondominated sorting genetic algorithm. Results demonstrate that the proposed solution achieves high levels of reliability as well as low latency.**

***Index Terms*— Capacitated reliable facility location problem, edge devices, 5G, NFV.**

## I. Introduction

Mobile operators have been witnessing an astonishing increase both in connectivity and in mobile data traffic [1]. It is expected that by 2021 there will be 11.6 billion mobile-connected devices, and a 7 fold increase in mobile traffic, driven mainly by video and social media [2]. Moreover, the trend towards more powerful mobile devices encourages new services, such as virtual reality (VR) and augmented reality (AR), multi-user interaction and 3D services. In such a scenario, localized live broadcast service will be common such as broadcasting of a live event (e.g., concert, sports event) and live broadcasting for social groups with ultra high definition or even with 3D and AR/VR technologies.

However, the large amount of bandwidth associated with these emergent applications poses several challenges to mobile network operators (MNO)s. A very dense coverage is required, with increased cell capacity to bring applications close to the end user and make the overall experience consistent. The evolution of technology towards the emerging fifth cellular technology (5G) is an attempt to meet these challenges. 5G networks should provide intelligent mechanisms to deliver services close to the end user in densely populated areas. Such mechanisms should perform in real time and in a cost-effective way. Moreover, 5G will employ heterogeneous architectures composed of small and macro cells connected to the core network by wired/wireless links [3].

To satisfy these requirements, mobile operators have deployed enhancements to their Long Term Evolution (LTE) networks for the future deployment of 5G networks. One of these enhancements is the LTE-broadcasting service as reported in 3GPP Releases 12 and 14 [4], [5], which has opened new opportunities for operators, and content providers to create innovative services and augmented broadcasting services. Moreover, 5G networks must support high data rates, with very low latency, as well as high reliability levels (reliability requirements of five nines 99.999%) [6].

The 5G network is expected to provide flexible and programmable capabilities [7]. For that, technologies such as network function virtualization (NFV) and software-defined networking (SDN) will be deployed, and their combination can facilitate the support of requirements of a variety of use cases [8], including multicast/broadcast video distribution. In a virtualized LTE-broadcasting network aiming at 5G, the main components of the LTE-broadcasting service will be implemented as virtualized network functions (VNFs), enhanced with both network-aware and content-aware adaptation capabilities. NFV has already been employed to implement mobile virtualized infrastructure [9]. One such case is described, in [10], where the authors have proposed a virtualized evolved packet core (EPC) as a service, allowing on-demand creation of mobile core networks. In another case detailed in [11], the authors proposed a multi-tenant virtualized infrastructure of small cells.

An ultra-dense 5G NFV-based network with computing devices at the edge of the network [12] is considered here; it furnishes local services to end-users in crowded areas. One common issue in the design of virtualized backhaul/fronthaul networks is the distribution of functionalities provided by a centralized cloud RAN to the geo-distributed edge devices [13]. These edge devices, also known as fog nodes, provide computing and storage capabilities for small cells and bring cloud functionalities to the edge of the network [14]–[16]. Edge devices are capable of hosting virtualized network functions (VNFs), have limited capacity

Manuscript received May 1, 2018; revised August 18, 2018; accepted August 28, 2018. Date of publication September 13, 2018; date of current version November 28, 2018. This work was supported in part by PEC-PG-CAPES/CNPq, Brazil, in part by the Sao Paulo Research Foundation under Grant #15/24494-8, in part by CNPq, in part by the European Union's Horizon 2020 Project under Grant 688941 (FUTEBOL), and in part by MCTI through RNP. *(Corresponding author: Nelson L. S. da Fonseca.)*

The authors are with the Institute of Computing, State University of Campinas, Campinas 13083-852, Brazil (e-mail: hernani@lrc.ic.unicamp.br; nfonseca@ic.unicamp.br).

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

Digital Object Identifier 10.1109/JSAC.2018.2869966

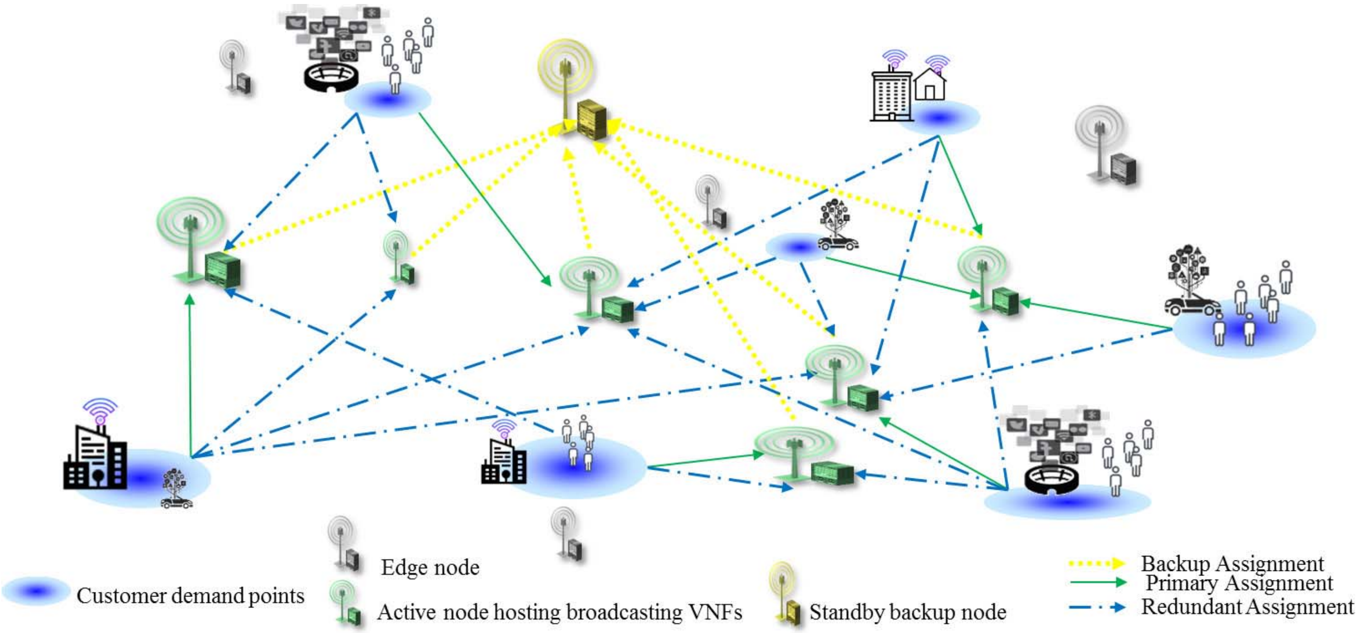

Fig. 1. Redundancy of edge devices.

and are prone to service outages, especially those arising from overloads from flash crowd event or the signaling storms resulting from core network function failures. The small cells with such edge devices are called cloud-enabled cells [3].

In the face of massive use of video services emerging today and especially broadcast services, a central question for MNOs is the location for the deployment of VNFs for broadcasting services, since the limited capacity of edge nodes and 5G requirements must be considered. The goal is to deploy, an optimized, reliable set of edge devices with zero perceived downtime for broadcast service provisioning while minimizing service costs, and service response time, as well as the loss of requests to establish a broadcast transmission. Indeed, the ability to provide massive video services is intimately related to the location where VNFs are deployed. In [17], the problem of network function placement is investigated with the aim of placing virtualized instances of Packet Data Network Gateways (PDN-GW) in a carrier cloud as well as for selecting the best gateway to promote traffic load balancing.

The problem of locating reliable edge devices can be formulated as a generalization of the classical facility location problem [18], otherwise known as a capacitated reliable facility location problem with failure probability (CRFLP) [19]. In the problem investigated in this paper, the facilities are edge devices which store the whole chain of VNFs implementing a broadcast transmission. Specifically, the edge device location problem has the following characteristics: i) customer demands initiated at a given demand point can be served by different facilities (multi-source); ii) the capacity of an active facility cannot be exceeded by the demand imposed by the demand points assigned to the facility (capacity-constraint); iii) each facility is instantiated with a given cost; iv) the facilities may fail from time to time according to a given failure probability.

In the CRFLP formulation proposed here, a standby backup facility will serve the demands hosted by another facility if the primary facility fails (Fig. 1), or if the load on that facility reaches its capacity. In case of failure, the backup facility will host the entire chain of VNFs once hosted by the primary facility which has failed. This backup facility is shared among $N$ other facilities, providing a $1 : N$ protection scheme. Moreover, if the designated backup facility is not available, $k$ other facilities can host the demand, thus enhancing the $1 : N$ protection scheme. These $k$ facilities are not reserved to provide a $1 : N$ protection scheme but can be employed to provide service on the fly in the absence of a backup facility.

The original contribution of this work is the formulation of the edge device location problem and the algorithms to solve it so that protection can be provided for 5G NFV-based small cells. Conflicting objectives are addressed by formulating a multi-objective optimization problem so that low service response time, high reliability, extensive coverage and low service provisioning cost are guaranteed. The proposed formulation is not restricted to the introduction of the massive video services. Indeed, the solution is valid for any service chain of VNFs hosted as a whole in a service facility.

The reliable edge device location problem is solved by employing two state-of-the-art Multi-Objective Evolutionary Algorithm (MOEA): an extended genetic NSGA-II algorithm [20], and an extended Particle Swarm Optimization MOPSO [21]. The hypervolume quality indicator was used to evaluate the quality of the solution provided by these two optimization algorithms.

This paper is structured as follows. Section II discusses related work. Section III presents the background material on LTE-Broadcasting services. Section IV presents the statement of the problem, while the problem formulation is introduced in Section  V. Section VI introduces metaheuristic to solve the target multi-objective optimization problem. Section VIII presents numerical results and Section IX concludes the paper.

## II. Related Work

In this section, some of the research work related to the virtual network function placement problem is presented, this can be seen as a generalization of the virtual network embedding problem [22] [23], a *NP*-hard problem [24]. Some authors approach the network function placement problem focusing only on QoS metrics while others also consider reliability. Basta *et al.* [25] propose an Integer Linear programming (ILP) model for optimal network function placement in the context of cellular networks, designed to minimize the network load while satisfying data-plane delay constraints.

In [26], a optimal placement of virtualized middlebox function is proposed to minimize the packet transmission cost in a virtualized Evolved Packet Core (vEPC). A vEPC joint embedding problem was proposed in [27], considering the latency of the VNFs. This problem is formulated as a Mixed Integer Linear Programming (MILP) one. Dietrich *et al.* [28] introduced a MILP formulation and use a relaxed Integer Linear Programming (ILP) to solve the network function placement problem, considering capacity constraints and the delay budgets between EPC components. The proposed ILP solution produces better load balancing, request acceptance rates, and resource utilization than does the MILP.

In [29], a coordinated approach for mapping and scheduling of VNFs to minimize different criteria such as cost, revenue, and service processing time is presented; it employs a greedy-based algorithm and a tabu search-based heuristic.

In [30], the problem of resource allocation in NFV-based networks using a binary search formulation is introduced for the minimization of the number of VNFs deployed. Both link transmission delays and VNF processing delays are considered. A delay aware resource optimization algorithm is proposed to achieve minimum end-to-end delays. However, such delays are assumed to be fixed values which is not a realistic assumption.

Ksentini *et al.* [31] formulated a VNF placement problem for the creation of virtual mobile networks. An algorithm to evaluate the trade-off between reducing the VNF relocation and the load on the Serving GateWay (VNF) was proposed, with the aim of reducing the latency of flow installation. In [32], it was proposed a solution for determining the optimal number of VNFs and their locations in a federated cloud under the demand imposed by the mobile traffic. Two algorithms were proposed: a MILP to derive the optimal number of VNFs, and a coalitional game to determine the optimal placement of VNFs in a federated cloud. In [33], optimal placement of virtual resources is proposed based on a canonical domain framework using the Schwartz-Christoffel conformal mapping in order to tackle the dynamic characteristics of mobile edge clouds. The proposed solution reduces the end-to-end latency as well as the total number of activated VMs.

In [34], a VNF placement algorithm for virtual 5G networks is proposed to minimize the length of the paths between users and their associated data anchor gateways as well as to optimize their sessions mobility.

Redundancy is an essential aspect of service provisioning, so that service resilience can be assured [35]. Service can be impacted by a failure of VNFs running on a virtual machine (VM), or by the failure of the physical machine hosting a VM [36].

Several papers in the literature have tackled different failure scenarios in network virtualization, including failure of the network nodes and links [37], [38]. In [39], a redundancy series-parallel model was introduced to improve the reliability of LTE eMBMS services. The model was formulated as a VNF redundancy allocation problem, and the trade-off between minimizing the processing delay of services and the required number of redundant VNF components was evaluated. Guerzoni *et al.* [40] formulated a MIP to address virtual network embedding under the constraint of reliability requirements for the physical substrate. In [41], a reliability-aware joint VNF placement is proposed in the context of NFV-enabled datacenter networks. The delay constraint is considered in the placement decision to maximize reliability and minimize the end-to-end delay network services. A MILP and a heuristic algorithm were proposed to solve the problem.

Ahuja and Krunz [42] studied a server placement problem formulated as a MILP problem with the goal of identifying the minimum number of servers needed to provide reliable streaming. Latency and reliability parameters were considered as input parameters.

Mangili *et al.* [43] proposed an optimal placement of VNFs content delivery network (CDN) replica servers to minimize the operational cost of virtual CDN nodes. To meet traffic demand resilience, the problem is formulated as a stochastic mixed integer programming problem. An approximate algorithm is designed to find the optimal placement of VNF CDNs. The work in [44] poses a problem for optimal placement of VNFs composing a CDN slice. The problem is formulated as two different ILPs that minimize costs and maximize the support of QoE requirements for virtual streaming services.

Furuta *et al.* [45] have proposed the placement of a distributed service by employing an uncapacitated facility location problem for selecting nodes to act as cluster heads for hierarchical routing. Laoutaris *et al.* [46] present a distributed algorithm to solve the uncapacitated $k$-median problem by focusing on the neighborhood of the nodes hosting a service instance.

Table I compares the VNF placement problems described in this paper in relation to QoS metrics, reliability and the evaluation method. Our work is the first to address the optimal placement of edge devices for reliable broadcasting service in 5G NFV-based small cells. We formulate a multi-objective optimization model which minimizes the costs of service provisioning, service processing time, and loss probability.

## III. LTE-Broadcasting Services

The LTE-Broadcasting service defined by 3GPP in Release (9) [47] (Fig. 2), also known as evolved multimedia broadcast multicast service (eMBMS), was designed

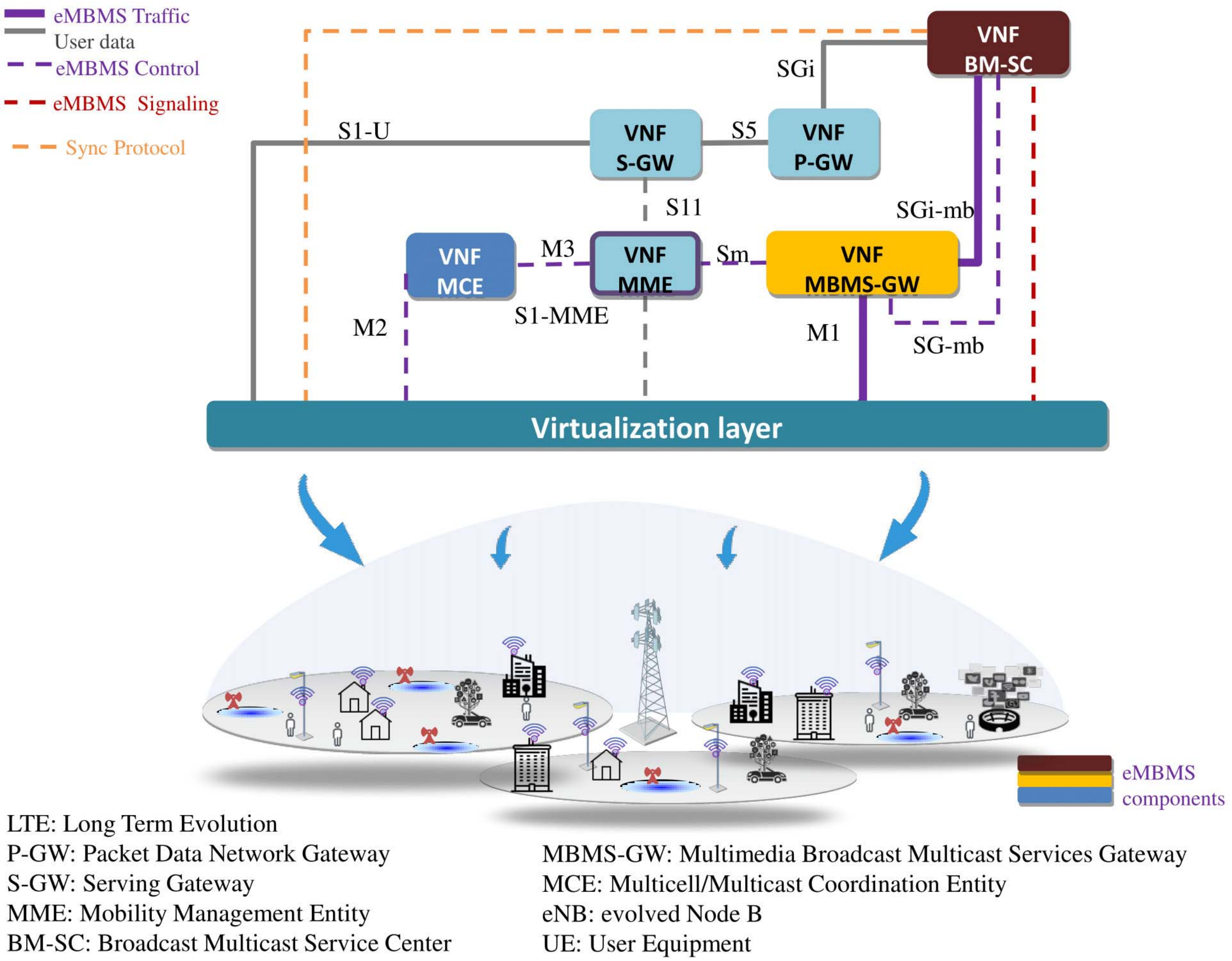

Fig. 2. LTE-broadcast in 5G NFV-based network.

TABLE I
COMPARISON OF ASPECTS COVERED BY RELATED PAPERS

| Approach | QoS metrics | Reliability | Multi-objective |
|---|---|---|---|
| [25] | Yes | No | No |
| [26] | Yes | No | No |
| [27] | Yes | No | No |
| [28] | Yes | No | No |
| [29] | Yes | No | Yes |
| [30] | Yes | No | No |
| [32] | Yes | No | Yes |
| [33] | Yes | No | Yes |
| [34] | Yes | No | Yes |
| [39] | Yes | Yes | No |
| [44] | Yes | Yes | No |
| [37] | Yes | Yes | No |
| [38] | Yes | Yes | No |
| [40] | Yes | Yes | No |
| [41] | Yes | Yes | No |
| [42] | Yes | Yes | No |
| [43] | Yes | Yes | No |
| [45] | Yes | Yes | No |
| [46] | Yes | Yes | No |
| our work | Yes | Yes | Yes |

to deliver media content to multiple devices in an LTE network. LTE Broadcast can be used for diverse use cases, namely, content distribution for connected vehicles, massive updates for IoT, critical information for Public Warning Systems (PWS), content pre-positioning. It supports all defined LTE bandwidths and formats, including FDD, TDD, and carrier aggregation. It was expanded in 3GPP Release 14 [5] with the introduction of Multicast/Broadcast single frequency network (SFN) and single cell point-to-multipoint for vehicular to everything (V2X), for enhanced machine-type communication (eMTC) and narrowband-IoT (NB-IoT), with the aim of supporting public broadcasting requirements [48]. This service introduces three main components into the LTE architecture: the Broadcast Multicast Service Center (BM-SC), the Multimedia Broadcast Multicast Services Gateway (MBMS-GW) components, and the Multicell/Multicast Coordination Entity (MCE).

The Broadcast Multicast Service Center (BM-SC), located at the evolved packet core, serves as an entry point for content providers to inject multimedia content into the LTE network. It also schedules and delivers MBMS transmissions. The Multimedia Broadcast Multicast Services Gateway (MBMS-GW) distributes IP multicast packets to all evolved Node B (eNBs) that are part of the eMBMS service. It performs MBMS session control signaling (session start/stop) to the E-UTRAN

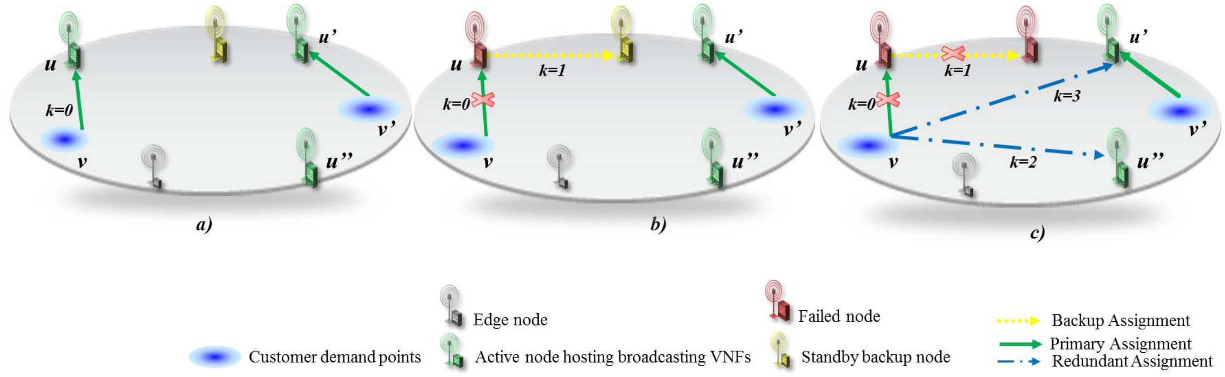

Fig. 3. Model of a protection scheme for the problem of capacitated reliable facility location with failure probability.

using the Sm interface with the Mobility Management Entity (MME), which coordinates signaling in MBMS sessions (session start/update/stop), delivers MBMS information to the MCE, including information on quality of service (QoS), using the M3 interface. The Multicell/Multicast Coordination Entity (MCE) component manages the radio resources for MBMS to all radios that are part of the MBSFN service area. It coordinates the transmission of synchronized signals from different cells (eNBs) using the M2 interface. The eNBs gather information about starting and stoping of session to be transmitted to mobile devices (UE). The Multimedia Broadcast Multicast Service Single Frequency Network (MBSFN) area consists of a group of synchronized radio cells, seen as a single transmission by a mobile device.

In 5G NFV-based networks, the LTE-Broadcast components are provided as virtual network functions, e.g., virtualized eMBMS components (e.g., vBM-SC, vMBMS-GW, vMME, vMCE) and hosted at various network points of presence located at small cell gateways to operate as aggregation points in order to improve signal strength [49]. This enables high throughput and excellent coverage [50], [16]. The LTE-broadcast supports sending a single stream of data to a large number of end-users with both fixed and mobile devices in a specific area.

LTE Broadcast is a key feature to be incorporated in 5G system networks since MNOs can employ cloud-native virtual networks to offer a different type of services (both live and non-real time services), location (venue-specific, local, regional, national) and QoS. In 5G networks, the LTE Broadcast technology shall support content distribution (e.g., videos, live streaming, software/firmware updates) to mobile devices and to a large number of IoT, V2X, and VR applications.

Therefore, an optimal broadcast service placement in 5G NFV-based networks with close to zero outages should incorporate efficient and flexible allocation of resources in a way to enable cost-efficient and reliable content distribution framework by employing a distributed edge RAN architecture [51].

## IV. Statement of the Problem

The service provided by 5G virtualized networks must meet strict QoS requirements. The low latency requirements can be met by placing devices at the edge of the network, thus, avoiding access to servers in distant clouds. However, these devices are subject to outages for a variety of reasons, such as node failure and overloads resulting from flash crowd events. In 5G networks, the continuity of service provisioning must be assured at the five nines reliability level, while guaranteeing strict latency requirements upon recovery from failure events. Consequently, an elaborated reliability strategy should be employed to provide this assurance. Various different strategies can be employed, but they differ in the number of redundant devices and thus imply different costs.

This paper provides a solution for the problem of the choice of the number of edge devices to deploy and their positioning in a way that supports the latency and reliability requirements at a minimal cost.

An enhanced 1:N protection scheme is proposed. In this scheme, an edge device can be assigned as backup device for other $N$ edge devices in a traditional 1:N protection scheme, yet it avoids the high cost of providing the 1:1 type of protection in which an exclusive backup device is reserved for each primary device. However, as the backup device is shared, it may not be available when a primary device fails; in this case, service will be discontinued. To decrease the possibility of discontinuation of service, additional backup devices can be assigned. These additional backup devices for a customer demand point will be limited in number, and their activation as backup devices follows a pre-defined order. If the main backup device fails, this progression of devices will be available to replace the main backup device. In this way, service will be discontinued only if the main backup device and all the other $k$ additional backup devices are unavailable.

Figure 3 illustrates the protection scheme used in this paper for $k = 3$. Figure 3$a$ shows the customer demand point $v$ served by the edge device $u$ and the customer demand point $v'$ served by the primary edge device $u'$. Figure 3$b$ shows how a standby backup node can replace the primary edge device $u$, while Figure 3$c$ shows how that service can also be provided by the secondary and third backup edge devices $u'$ and $u''$.

The problem presented here is how to choose the number of edge devices to deploy and where to position them in a way to support the latency and reliability requirements of 5G services at a minimal cost so that edge devices are protected by a 1:N protection scheme enhanced by the possibility of using $k$ additional backup devices in the case of the failure of a shared backup device.

The problem assumes that customer demand points request video broadcast services with functionalities implemented by a VNF chain. All VNFs of that chain are hosted by the same primary edge device. In the case of failure, the whole chain is transferred to a single backup device. Service can be denied if any of the resources to host the VNF chain are unavailable; moreover it can be discontinued if no backup device is available to host the VNF chain.

## V. Problem Formulation

The reliable edge node location problem formulation and its objective functions are defined in this section.

Let $G = (U \cup V,\ E)$ be a bipartite graph in which $U$ denotes the set of edge devices, and $V$ defines the crowded areas. $U$ defines the set of potential edge nodes where the VNFs of broadcast services can be placed (activated) to serve the customer demand points $v \in V$ in the crowded area. The broadcast communication link between the nodes and demand points are defined by $E \subseteq U \times V$.

The model is formulated as a two-stage stochastic optimization program. In the first stage, the subset of nodes on which the VNFs services should be placed, designated active nodes are chosen. The standby backup nodes to provide redundancy in case of an active node failure are selected. In the second stage, demand points are optimally assigned to these activated nodes. The aim is to assign each customer to a primary node that will provide service under normal circumstances. If this node fails, or the utilization of its capacity reaches its limit, a standby backup facility will serve the demands hosted by the node in failure. A backup node is shared among $N$ other nodes, providing a $1 : N$ protection scheme. However, if the designated backup node is unavailable, $k$ other active nodes can host customer demands if they have capacity available. A $k$-level of redundancy means that there are $k$ active nodes with potential capacity available to host customer requests if the standby backup facility fails. The $1 : N$ protection scheme is enhanced by the $k$-level of redundancy, thus providing greater protection than a traditional $1 : N$ scheme. As output, the proposed model returns the optimal placement of the broadcast service.

Table II shows the notation used in the formulation proposed. The binary decision variables of the model are $y_u$ and $x_{uvk}$, $y_u$ denotes whether a node is active or not.

$$y_u = \begin{cases} 1 & \text{if node } u \text{ is active,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The binary decision variable $x_{uvk}$ represents the assignment of customer $v$ to the node $u$.

$$x_{uvk} = \begin{cases} 1 & \text{if node } u \text{ is the primary, backup or } k\text{-level} \\ & \text{redundant node to customer } v, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The nodes are ordered as follow: a primary node at level $k = 0$, a standby backup node at level $k = 1$, and $k$ other active nodes to enhance the $1 : N$ protection scheme at level 2 to $k + 1$.

TABLE II
Table of Notations

| Symbol | Description |
|---|---|
| $U$ | Set of potential edge nodes |
| $V$ | Set of customers (demand points) |
| $f_u$ | Cost to activate node $u \in U$ in \$ |
| $d_v$ | Amount of broadcast data selected by customer $v \in V$ |
| $c_{uv}$ | Bandwidth demand between a node $u$ and a demand point $v$ in Mbps |
| $k$ | Redundancy level at which a customer $v$ is assigned to a node $u$ |
| $l_{uv}$ | Communication latency between $u$ and $v$ in ms |
| $L$ | Service latency requirement in ms |
| $b_{uv}$ | Bandwidth capacity of the link between nodes $u$ and $v$ in Mbps |
| $C_u$ | Processing capacity of node $u$ in VM units |
| $n_{uv}$ | Processing request for demand point $v$ assigned to a node $u$ in VM units |
| $q_u$ | Failure probability of node $u$ |

Assume that the probability of failure of each node $u$ is denoted by $q_u$. A customer $v$ is assigned to an active node $u$, as $v$'s $k$ redundant node with the probability $(1 - q_u)$, if all lower level $(l < k)$ nodes failed with the probability $q_u$.

The goal of the proposed optimization model is to choose the locations of nodes meeting reliability and capacity constraints while minimizing overall service provisioning cost, service response time and service loss probability. In this optimization model, the objectives are formulated as follows:

*1) Provisioning Cost:* This is the total cost to activate the nodes from which the service will be provided (in \$). The overall service provisioning cost can be expressed as follows:

$$\sum_{u \in U} f_u y_u \tag{3}$$

where, $f_u$ denotes the price of activating a node *u*;

*2) Response Time:* It is the communication latency between an active node and a demand point. The response time is the response time from an active node or the response time from a standby backup node that will cover the demands from a failed node, or from any of $k$-redundant nodes.

The service response time can be formulated as follow:

$$\sum_{v \in V} \sum_{u \in U} \sum_{t=0}^{k+1} d_v (c_{uv})^{-1} x_{uvt} (1 - q_u) {q_u}^t \tag{4}$$

*3) Loss Probability:* This defines the probability of denying service to a service request of a demand point due to lack of resources. It is defined by the ratio between the number of requests not served and the total number of requests. The loss probability can be expressed as follows:

$$\frac{1}{|V|} \sum_{v \in V} \sum_{u \in U} (x_{uv0} q_u + \sum_{t=1}^{k+1} x_{uvt} {q_u}^t) \tag{5}$$

Achieving simultaneously the objectives defined above is not possible since they are clearly conflicting. In the next section, a multi-objective optimization problem is formulated so that the trade-off between these objectives can be further analyzed.

## VI. Multi-Objective Optimization Problem Formulation

This section introduces formulation for the multi-objective optimization problem formulation (MOOP) for solving the capacitated reliable facility location problem with failure probability version of the edge device location problem. In general, in a MOOP, the objective functions conflict with each other, producing a set of unique solutions, resulting from the consideration of conflicting trade-offs between the objectives. The edge device location problem is formulated as a MOOP, with three conflicting objectives: minimization of service provisioning cost, minimization of service response time and minimization of the loss probability.

We formulate the MOOP as follow:

$$\text{Min} \sum_{u\in U} f_u y_u \tag{6}$$

$$\text{Min} \sum_{v\in V}\sum_{u\in U}\sum_{t=0}^{k+1} d_v (c_{uv})^{-1} x_{uvt}(1-q_u){q_u}^t \tag{7}$$

$$\text{Min} \frac{1}{|V|}\sum_{v\in V}\sum_{u\in U}(x_{uv0}q_u + \sum_{t=1}^{k+1} x_{uvt}{q_u}^t) \tag{8}$$

$$\text{subject to: } \sum_{k=1}^{|U|} x_{uvk} \le 1 \quad \forall u \in U,\ v \in V, \tag{9}$$

$$\sum_{v\in V}\sum_{k=1}^{|U|} c_{uv}x_{uvk} \le b_{uv} \quad \forall u \in U | y_u = 1 \tag{10}$$

$$\sum_{v\in V}\sum_{k=1}^{|U|} l_{uv}x_{uvk} \le L \quad \forall u \in U | y_u = 1 \tag{11}$$

$$\sum_{v\in V}\sum_{k=1}^{|U|} n_{uv}x_{uvk} \le C_u \quad \forall u \in U | y_u = 1 \tag{12}$$

$$x_{uvk} \le y_u \quad \forall u \in U,\ v \in V,\ k = 1, \ldots, |U| \tag{13}$$

$$x_{uvk}, y_v \in \{0, 1\} \quad \forall u \in U,\ v \in V,\ k = 1, \ldots, |U| \tag{14}$$

The objective function defined by Equations (6), (7) and (8) minimizes the service provisioning cost, the service response time and the loss probability.

Constraint (9) prohibits a demand point from being allocated to a given node at more than one level. Constraint (10) ensures that the bandwidth allocated does not exceed the capacity of the link between a node $u$ and a demand point $v$. Constraint (11) guarantees that the latency between a node $u$ and a demand point $v$ does not exceed predefined service latency requirements. Constraint (12) ensures that the processing capacity of an active node is greater than or equal to the processing demand request of all demand points. Constraint (13) prohibits the allocation of a customer (demand points) on an inactive node. Constraint (14) expresses the domain of the binary decision variables.

The optimization problem is NP-hard, since it is a generalization of capacitated facility location with failure probability, and its computational complexity ($2^{|U|}$) is a function of the number of nodes. In the next section, two multi-objective meta-heuristic algorithms are introduced to solve the multi-objective optimization problem proposed in this section.

## VII. Evaluated Algorithms

This section introduces the nondominated sorting genetic algorithm (NSGA-II) and the multi-objective particle swarm optimization (MOPSO), proposed to solve the edge device location problem. The output of the two algorithms are Pareto fronts. A feasible solution $\vec{X}_1$ is non-dominated (Pareto optimal) if there is no other feasible solution $\vec{X}_2$ that dominates it. The set of non-dominated solutions $F(\vec{X}_1)$ in the objective space is known as Pareto front [52]. The Pareto fronts are generated from all the fronts of the solutions produced by the MOPSO and NSGA-II algorithms in each independent run.

### *A. Multi-Objective Genetic Algorithm*

The nondominated sorting genetic algorithm (NSGA-II) [20] is a population-based algorithm based on the theory of the natural evolution of individuals. An individual represents a point in a search space, i.e., a suggested solution for the edge device location problem. A nondominated sorting technique and genetics operators (selection, crossover, and mutation) are used to obtain best solutions and create a new population. In genetic algorithms (GA), over time, the individuals in the population will adapt to their environment through a process of evolution. The adaptation process evaluates the objective functions defined by the Equations (6), (7) and (8), and these are passed on to the next generation for solving the problem. Each generation consists of a population of multiple chromosomes, representing possible solution of the reliable edge device location problem. A chromosome is composed of several genes, representing the decision variables defined in Equations (1) and (2) denoting if a node $u$ is selected to host the VNFs of a broadcast service and what level of redundancy is, respectively. These genes or decision variables are then evaluated by given preference for better individuals. For each iteration, the quality of a chromosome or the solution is evaluated based on the value of the objective function, by the genetic operators. Algorithm 1 shows the pseudo-code of NSGA-II. The first step is to generate randomly an initial population, selecting the activated edge devices of each solution. For each solution of a population, the edge devices are selected randomly, and the solution is compared with every other solution in the emerging population to ensure the uniqueness of the solution. Unique solutions are added to the emerging population, and the non-unique solutions are rejected, a new solution is then generated. This process stops once $pSize$ distinct solutions have been reached. Once the solutions have been generated, the genetics operators are employed to create the next generation. Again, the best solutions from one generation to the next are chosen, and these randomly produce a new solution for each generation, such that each new solution is different from any of the solutions already in the emerging population.

**Algorithm 1:** Pseudo Code for NSGA-II

```
Input: $|U|$,$|V|$: $f_u$, $d_v$, $c_{uv}$, $C_u$, $q_u$, $k$, $b_{uv}$, $l_{uv}$,$L$,$pSize$,
       $iMax$.
/* $pSize$ population size */
/* $iMax$ max iteration */
Output: The optimal solutions $O_f$, $x_{uvk}$, $y_u$
begin
    $P_{(0)} \longleftarrow InitializePopulation()$ /* initialize
    population with size pSize */
    $ActivateNodes(U)$
    $AssignNodestoDemands(U, V)$/* assign
    nodes to demand points */
    $EvaluateFunctions(Eq.(6), Eq.(7), Eq.(8))$
    /* evaluate the objective functions
    */
    $CheckConstraints()$/* check constraints
    given by formulations (9 - 14). */
    $NondominatedSort(P_{(0)})$ /* the population
    $P_{(0)}$ is sorted in fronts */
    $Q_{(0)} \longleftarrow$
    $SelectandReproductionandImmigration(P_{(0)})$
    /* solution is evaluated through the
    genetics operators and another
    population is generated */
    $EvaluateFunctions(Eq.(6), Eq.(7), Eq.(8)) \longleftarrow$
    $Q_{(0)}$/* evaluate the objective
    functions */
    for $i = 1$ to pSize do
        $P_{(i+1)} \longleftarrow Merge(P_{(i)} \cup Q_{(i)})$/* */
        $EvaluateFunctions(Eq.(6), Eq.(7), Eq.(8))\longleftarrow$
        $P_{(i+1)}$/* *//* evaluate the
        objective functions */
        $CheckConstraints()$/* check
        constraints given by
        formulations (9 - 14). */
        $NondominatedSort(P_{(i+1)})$ /* the
        population is sorted in fronts */
    end
    while $t < iMax$ do
        $P_t \longleftarrow NondominatedSort(P_i)$/* select
        parent from population */
        $P_t^{'} \longleftarrow SBX(P_{t1}, P_{t2})$/* perform
        crossover */
        $P_t^{'} \longleftarrow Mutation(P_t)$/* perform mutation
        */
        $EvaluateFunctions(Eq.(6), Eq.(7), Eq.(8)) \longleftarrow$
        $P_t^{'}$/* evaluate children */
        $CheckConstraints()$/* check
        constraints given by
        formulations (9 - 14). */
        $P \longleftarrow NondominatedSort(P_t^{'})$ /* get the
        nondominated solution */
        $t$= $t$+2
    end
    $O_f \longleftarrow P$ /* return the nondominated
    fronts of the solution set */
end
```

Then, through the mutation process, the genes or decision variables are randomly changed, introducing randomness to each generation. The crossover operator is used for mating different chromosomes, generating a random solution with probability $p_c$. The solution selected mutates with probability $p_m$, producing a better solution to be added to the new population set. The process of adding solutions to the population continues till the population contains a total of $pSize$ solutions.

In the proposed NSGA-II algorithm, a limiting value called max generation is used as a termination criterion to decide when the algorithm should stop. The computational complexity of the NSGA-II is $O(Mn^2)$, where $M$ is the number of objective functions and $n$ is the population size [20].

### *B. Multi-Objective Particle Swarm Optimization*

Multi-Objective Particle Swarm Optimization (MOPSO) is a population-based meta-heuristic inspired by nature, whereas instead of evolution, it is built on social interactions among members in schooling fish or flocking birds. The population, i.e., the selected nodes to host the VNFs of a broadcast service and its level of redundancy, is called a swarm, and each unique solution of the multi-objective edge device location problem is referred to as a particle. A particle represents the set of binary variables representing the solution of the edge device location problem namely: $y_u$, $x_{uvk}$. Every $i$-th particle or individual solution is associated with a position in the search space $\vec{X}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$ and a velocity attribute $\vec{V}_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,d})$ associated with its movement in a $d$-dimensional search space.

$$v_{i,j} = wv_{i,j} + c_1\rho_1[pb_{i,j} - x_{i,j}] + c_2\rho_2[pb_{g,j} - x_{i,j}] \quad (15)$$

$$x_{i,j} = x_{i,j} + v_{i,j} \quad (16)$$

To find the optimal solution, the particle or the possible solution of the edge device location problem updates its movements according to Equations (15) and (16), representing velocity and position respectively. The constants $c_1$ and $c_2$ represent cognitive learning and social learning, respectively, $\rho_1$ and $\rho_2$ are uniformly-distributed random numbers in [0,1]. The coefficient $w$ denotes the inertia used to prevent the unbounded growth of the particle velocities.

During the search process, whenever the best solution (e.g., set of edge devices with low provisioning cost, low response time and low loss probability) is found for each particle (local best) $pb_{i,j}$, a new leader, i.e a new best particle (the best solution of the edge device location problem) among the swarm (global best) $pb_{g,j}$, is selected to guide the other particles towards the best regions in the search space.

The MOPSO outputs a set of unique solutions, known as a Pareto-optimal set, resulting from the trade-offs between the objectives defined by Equations (6), (7) and (8) aiming at minimizing the service provisioning cost, the service response time and the loss probability, respectively.

Algorithm 2 presents the pseudo-code of the MOPSO algorithm. The first step is the random generation of an initial population, activating a set of nodes and assigning the customer demand point to the nodes with a level of

**Algorithm 2:** Pseudo Code for MOPSO

```
Input: $|U|$,$|V|$: $f_u$, $d_v$, $c_{uv}$, $C_u$, $q_u$, $k$, $b_{uv}$, $l_{uv}$,$L$,$pSize$, $iMax$.
/* $pSize$ swarm size */
/* $iMax$ max iteration */
/* $aSize$ archive size */
Output: The optimal solutions $O_f$, $x_{uvk}$, $y_u$
begin
    $ActivateNodes(U)$
    $S_0 \longleftarrow AssignNodestoDemands(U, V)$ /* initialize the nondominated solutions from population $S_0$ */
    $EvaluateFunctions(Eq.(6), Eq.(7), Eq.(8))$ /* evaluate the objective functions */
    $CheckConstraints()$ /* check constraints given by formulations (9 - 14). */
    $v_{i,j} \longleftarrow \emptyset$ /* initialize the speed of each particle */
    $pb_{i,j} \longleftarrow x_{i,j}$ /* initialize the position of each particle */
    $A_0 \longleftarrow nondominated(S_0)$ /* initialize the archive */
    $CrowdDistance(S_0)$ /* ensure the diversity of the solutions */
    for $t = 1$ to $iMax$ do
        for $i = 1$ to $pSize$ do
            $EvaluateFunctions(Eq.(6), Eq.(7), Eq.(8))$ /* evaluate the objective functions */
            $CheckConstraints()$ /* check constraints given by formulations (9 - 14). */
            $A_t \longleftarrow nondominated(S_t \cup A_t)$ /* update the archive */
            $Gb \longleftarrow get_gbest()$ /* return global best */
            $pb_i \longleftarrow get_pbest()$ /* return local best */
            $v_{i,j} = wv_{i,j} + c_1\rho_1[pb_{i,j} - x_{i,j}] + c_2\rho_2[pb_{g,j} - x_{i,j}]$ /* compute the speed of each particle */
            $x_{i,j} = x_{i,j} + v_{i,j}$ /* compute new position of each particle */
        end
        if $l_t > aSize$ then /* $l_t$ size of the archive */
            $TruncateArchive()$
        end
        $CrowdDistance(S_t)$ /* ensure the diversity of the solutions */
    end
    $O_f \longleftarrow A_t$ /* return the results */
end
```

redundancy. Furthermore, the constraint satisfaction is checked by Equations 9 to 14 and the objective functions defined by the Equations (6), (7) and (8) are evaluated for every generation of the swarm in order to provide different nondominated solutions.

Every particle or unique solution of the edge device location problem is compared with the global best (e.g., set of edge devices with low provisioning cost, low response time and low loss probability) from the archive. To ensure the diversity of the nondominated solutions, a crowded distance approach [53] is adopted. The crowded distance technique estimates the density of solutions, calculating the average distance of two points on either side of these points along each of the objectives [20]. The MOPSO maintains an external archive to keep historical records of the nondominated solutions found along the search process.

The movement Equations (15) and (16) are iterated, until a max iteration is reached. The computational complexity of the MOPSO is $O(Mnlogn)$ ($M$ is the number of objective functions to be optimized, and $n$ is the population size) [54].

To overcome the limitations of the MOEA related to diversity and convergence of the obtained nondominated solution set, the hypervolume (HV) quality indicator for performance metric [55] was used for the evaluation of the algorithms proposed.

## VIII. Performance Evaluation

In this section, we show the trade-off between the conflicting objectives in the solution of the reliable edge device locating problem.

*1) Evaluation Environment:* The MOEA algorithms run on JMetal version 5.3, a Java-based framework for multi-objective optimization [56] and on a Debian GNU/Linux Squeeze, with two Intel Xeon (2.13GHz) with 4 cores each, and 78GB RAM.

The simulated scenario for providing live broadcast service to demand points is organized in service areas (cells), e.g., area in which data of a specific MBMS session are sent. The areas can be outdoors or indoors edge device in a venue. In the scenario explored here, the MNO provides access to a distributed infrastructure with cloud-enabled cells placed on the network edge [57] with LTE-Broadcast capabilities for provisioning live broadcast service to crowded areas. It is considered that users data rate at the cell edge is uniformly distributed in the interval [100, 200]Mbps [58]. The data rate capacity supported by small cells are uniformly distributed in the range [2, 9]Gbps [58]. The density of small cells to be deployed is 10 to 70 BSs/$km^2$. An ultra-dense 5G network scenario is assumed with dimensions of 0.4$km \times 0.4km$ [59]. Following the work in [60], the cost of activation of the nodes if uniformly distributed in [$0.03499, $0.0698].

Different redundancy levels $k = \{3; 5; 15\}$ were analyzed. For the sake of simplicity, all nodes were considered to have the same failure probability, i.e., $q_u = q \; \forall u \in U$. A customer is assigned to an active node with the following probability $(1-q) = \{0.9; 0.9999; 0.99999\}$. The protection scheme $1 : N$ was evaluated for $N = 30$ nodes.

*2) Parameterization:* For the comparison of the proposed MOEA, the following parameters were chosen. The population size and the swarm size for the NSGA-II and MOPSO were set to 100. In the MOPSO, a combination of uniform and non-uniform mutation was employed. The size of the external

TABLE III
PARAMETERITION OF THE MOEA [20]

| | |
|---|---|
| | NSGA-II [20] |
| Population Size | $pSize = 100$ individuals |
| Crossover | distribution index $\eta_c = 20$ |
| Crossover probability | $p_c = 0.9$ |
| Polymonial mutation | distribution index $\eta_m = 20$ |
| Mutation Probability | $p_m = 1/L$ where $L$ is the number of decision variables |
| | MOPSO [21] |
| Swarm Size | $sSize = 100$ individuals |
| Mutation | uniform + non-uniform |
| Archive size | $aSize = 100$ individuals |
| Cognitive learning | $c_1 = [1.5, 2.0]$ |
| Social learning | $c_2 = [1.5, 2.0]$ |
| Inertia | $w = [0.1, 0.5]$ |
| | $\rho_1$ and $\rho_2$ uniformly-distributed in [0,1] |
| 25 independent run.<br>Maxiteration 1 000<br>25 000 evaluation function | |

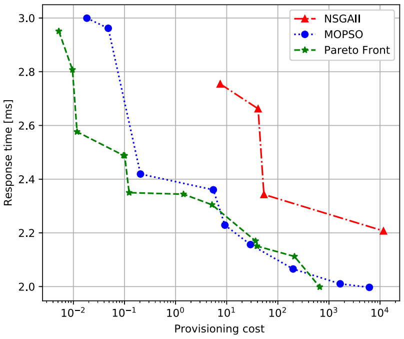

Fig. 4. Pareto front of provisioning cost vs. response time.

archive used to store the nondominated solutions of MOPSO is 100. For the NSGA-II, a simulated binary crossover (SBX) operator and polynomial mutation were used with distribution indices of $\eta_c = 20$ and $\eta_m = 20$, respectively [20]. The crossover for mating has a probability $p_c = 0.9$ and the mutation probabilities $p_m = 1/L$, where $L$ is the number of decision variables involved in making the mutation. Table III summarizes the MOEA parameters settings used.

*3) Results:* In this subsection, results obtained by employing the proposed algorithms are discussed and the performance of the algorithms to compute the approximated nondominated solution for the Pareto fronts evaluated. A maximum of 25,000 function evaluations was run. Approximated Pareto fronts were obtained after 1,000 iterations. For each experiment, 25 independent runs were performed, a confidence level of 95% was considered.

A representative set of results will be presented in this section. Figures. 4, 5 and 6 depict the nondominated solutions obtained for an experiment with the probability for a customer (demand points) being successfully assigned to a node $(1-q) = 0.9999$ and level of redundancy of $k = 5$.

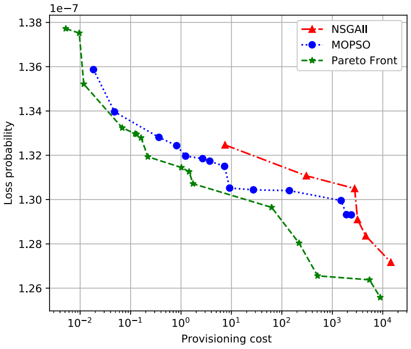

Fig. 5. Pareto front of provisioning cost vs. loss probability.

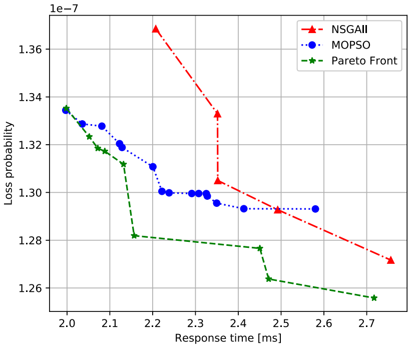

Fig. 6. Pareto front of response time vs. loss probability.

Figure 4 illustrates the trade-off between service provisioning cost and service response time. Results show that the service response time decreases with the increase in service provisioning cost. Figures 5 and 6 show the results for the loss probability. It can be seen that the loss probability decreases as both service provisioning cost and response time increase. Provisioning cost increases since a larger number of nodes are active. Reducing the number of active nodes can lead to the creation of resource bottlenecks and, consequently, an increase in the service response time, as well as in the loss probability. Consequently, activating more nodes increases redundancy and enhances reliability, although increasing the service provisioning cost. In Figures 5 and 6, the loss probability was in the order of $10^{-7}$ and the service response time was in the range $[1.9, 2.6]\ ms$.

The results depicted in Figures. 4, 5 and 6 show that the MOPSO provides faster convergence to the Pareto front than the NSGA-II algorithm. It was concluded that the crowding distance approach used in the MOPSO was helpful in providing both convergence and diversity of nondominated solutions closest to the Pareto fronts.

To demonstrate how the level of redundancy influences on distinct objective functions, the redundancy level was varied for: $k = 3$, $k = 5$, $k = 15$ ( Figs. 7, 8, and 9). As expected, the greater the redundancy level, the lower the loss probability, because the number of the possible reallocation of the demand of failed node is greater. These results clearly indicate the advantage of adopting the enhanced protection scheme created by the redundancy in addition to the traditional $1 : N$ protection scheme. Selecting the right configuration should also consider the impact of backup redundancy level on the service response time and provisioning cost. If on the one hand, high-levels of redundancy result in low loss probability (Fig. 8), on the other hand, they result in high service response

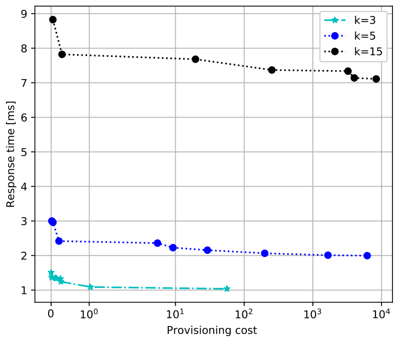

Fig. 7. Provisioning cost vs. response time for different redundancy levels.

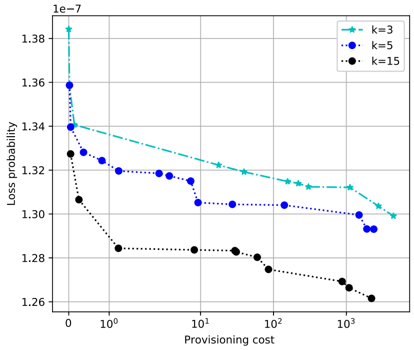

Fig. 8. Provisioning cost vs. loss probability for different redundancy levels.

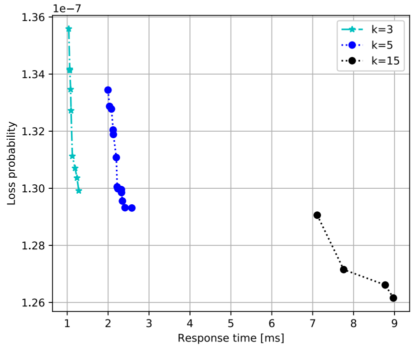

Fig. 9. Response time vs. loss probability for different redundancy levels.

time (Fig. 7). A high level of redundancy means a higher number of active nodes, therefore more information must be processed, resulting in additional response time. It shows that when $k = 3$ the service provisioning cost was lower than when $k = 5$ and $k = 15$ were adopted. This may be related to the fact that fewer nodes were activated. As a consequence, the provisioning cost was lower than that of with a high-level of redundancy.

Figure 10 depicts the trade-off between service response time and service provisioning cost. Results show that the service response time decreases with an increase in the provisioning cost. This is due to the greater availability of resources when a larger number of nodes are available, consequently serving demands faster. Fig. 10 also shows that when the probability of a customer (demand point) being successfully assigned to a node is low (e.g., $(1 - q) = 0.9$) the service response time is high. This is due to the reconfiguration required to reestablish the broadcast sessions from failed nodes. In Figures 11 to 14, it can be seen for $(1-q) = 0.99999$ and $(1-q) = 0.9$ the loss probability was in the order of $10^{-9}$ and $10^{-1}$, respectively.

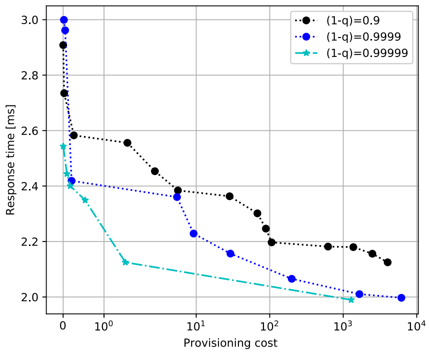

Fig. 10. Provisioning cost vs. response time for different failure probability.

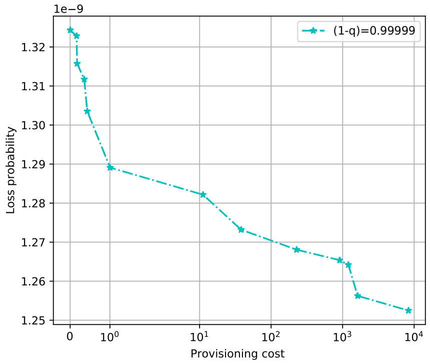

Fig. 11. Provisioning cost vs. loss probability.

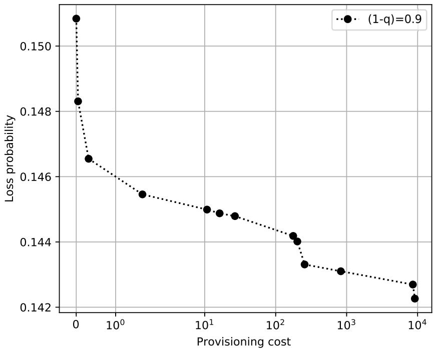

Fig. 12. Provisioning cost vs. loss probability.

Figures 15 and 16 show the impact of different latency constraints on the service response time, loss probability and the provisioning cost. Figure 15 shows that for a latency constraint of 10 ms, the loss probability was slightly lower than when the latency constraint was 1 ms. This is due to the distance between the serving nodes and the customer demand points. Figure 15 also shows that when latency constraint of 1 ms, the provision cost was greater than that when the latency constraint was 10 ms. This result indicates that the model employed activated more nodes to compensate for the loss

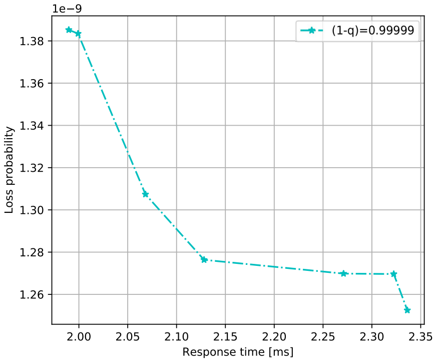

Fig. 13. Response time vs. loss probability.

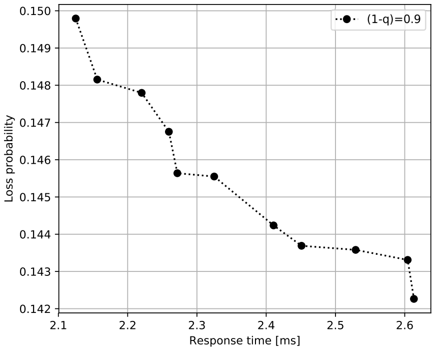

Fig. 14. Response time vs. loss probability.

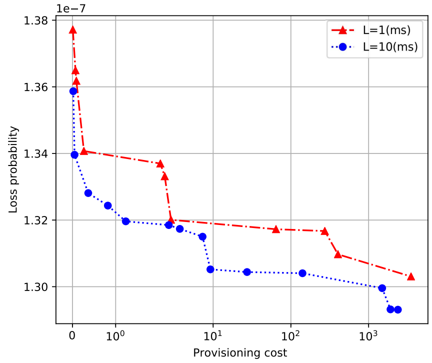

Fig. 15. Provisioning cost vs. loss probability.

probability experienced under a 1 ms latency constraint. Figure 16 shows that with a 10 ms latency constraint the service response time is slightly higher than 0.2 when compared to that when the latency constraint is 1 ms. This result is reasonable, as with a 10 ms latency constraint, the model is likely to have a certain latency budget, making it possible to afford the assignment of a customer to a relatively distant node.

Figures 17 and 18 show the influence of the number of nodes on the objective function in a scenario with a probability $(1-q) = 0.9999$ and a backup redundancy level of $k = 5$. Fig. 17 shows that the resulting loss probability was in the order of $10^{-4}$. The loss probability increases slightly when the number of available nodes increases. This is due to the fact that fewer nodes were employed in the redundancy model.

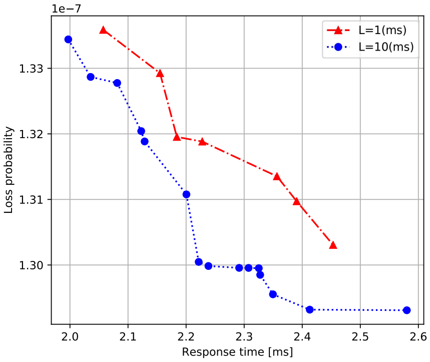

Fig. 16. Response time vs. loss probability.

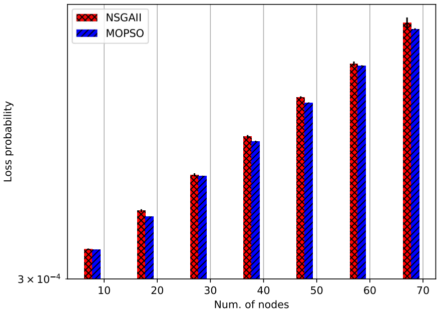

Fig. 17. Loss probability vs. number of nodes.

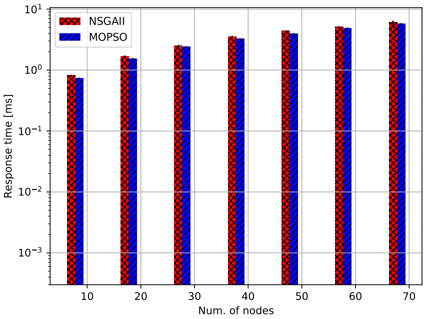

Fig. 18. Response time vs. number of nodes.

Figure 18 shows that the service response time increases with the number of nodes. The tradeoff between service provisioning cost vs. loss probability can be interpreted as follow: To minimize the provisioning cost, the number of active nodes must be reduced, but fewer resources will then be available to serve demands, this reduction will be accompanied by an increase in loss probability.

In summary, furnishing resilience and fault-tolerance can be costly, as a significant number of redundant nodes are required to meet the expected strict reliability requirements. These experiments have demonstrated that when the $1:N$ protection scheme is enhanced by a $k$-level of redundancy, higher levels of reliability and lower latency can be achieved.

The quality of the results given by the proposed algorithms is analyzed, by employing the hypervolume quality indicator (HV), which measures both convergence and diversity of the approximated Pareto front. Table IV includes the values of

TABLE IV
HV. MEDIAN AND INTERQUARTILE RANGE

| | NSGAII | MOPSO |
|---|---|---|
| CRFLP | $2.96e-01_{4.2e-02}$ | $6.95e-01_{2.6e-02}$ |

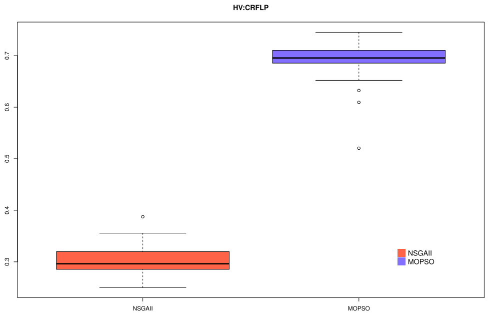

Fig. 19. Approximated solution using HV quality indicator with NSGA-II and MOPSO.

the median and the interquartile range in the measurements of the location and the statistical dispersion of the approximated Pareto fronts obtained (nondominated solutions), respectively. Table IV shows that the MOPSO algorithm converged faster than did the NSGA-II algorithm, producing approximated nondominated solutions closer to the Pareto front. The boxplot depicted in Fig. 19 illustrates the values of the HV indicators used to assess the quality of the resulting approximated nondominated solutions. Results show that the MOPSO algorithm converges faster to the Pareto front and diversified the solution set obtained more than did the NSGA-II algorithm.

## IX. CONCLUSION

In this paper, the problem of furnishing reliable broadcast services in a 5G NFV-based small cell network was investigated. The chain of VNF functions is stored in edge devices closer to the end user. The main question addressed was where to locate these edge devices to provide reliable broadcast services. The problem was formulated as a capacitated reliable facility location with failure probability. A multi-objective optimization approach was proposed, focusing on the location of the edge nodes from which the virtualized broadcasting service would be activated so that the service response time, service loss probability and service provisioning cost were minimized. A low service response time and high-level of reliability were obtained at the cost of a large number of redundant active nodes, thus increasing the service provisioning cost.

The proposed model was evaluated by employing multi-objective evolutionary algorithm (MOEAs), the NSGA-II and MOPSO, to compute the approximated Pareto fronts. Results showed that the MOPSO algorithm performs better than does the NSGA-II when both convergence and diversity of the obtained solution set to the Pareto front are considered.

The proposed formulation can be used to solve the edge device location problem for specific scenarios, including specific mobility models. Numerous scenarios were randomly generated, by the scattering of demand points to provide a first approximation of the position at mobile nodes.

## REFERENCES

[1] NGMN. (Feb. 2015). *NGMN 5G White Paper*. [Online]. Available: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN5WhitePaperV10.pdf

[2] Cisco. *Forecast and Methodology 2016–2021*. Accessed: Jan. 2018. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[3] I. Giannoulakis *et al.*, "The emergence of operator-neutral small cells as a strong case for cloud computing at the mobile edge," *Trans. Emerg. Telecommun. Technol.*, vol. 27, no. 9, pp. 1152–1159, 2016.

[4] *Multimedia Broadcast/Multicast Service (MBMS); Protocols and Codecs*, document TS 26.346, 3GPP, Oct. 2014.

[5] *LTE; Service Requirements for V2X Services (Release 14)*, document TS 122.185 v14.3.0, ETSI, 2017.

[6] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[7] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, and D. Darche, "On the scalability of 5G core network: The AMF case," in *Proc. 15th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2018, pp. 1–6.

[8] 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.

[9] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Netw.*, vol. 28, no. 6, pp. 18–26, Nov. 2014.

[10] T. Taleb *et al.*, "EASE: EPC as a service to ease mobile core network deployment over cloud," *IEEE Netw. Mag.*, vol. 29, no. 2, pp. 78–88, Mar. 2015.

[11] J. O. Fajardo *et al.*, "Introducing mobile edge computing capabilities through distributed 5G cloud enabled small cells," *Mobile Netw. Appl.*, vol. 21, no. 4, pp. 564–574, 2016.

[12] *ETSI Industry Specification Group Mobile-Edge Computing*, document ETSI GR MEC 017 V1.1.1, ETSI, Feb. 2018.

[13] 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.

[14] G. Premsankar, M. Di Francesco, and T. Taleb, "Edge computing for the Internet of Things: A case study," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1275–1284, Apr. 2018.

[15] F. Messaoudi, A. Ksentini, and P. Bertin, "On using edge computing for computation offloading in mobile network," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–7.

[16] 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.

[17] M. Bagaa, T. Taleb, and A. Ksentini, "Service-aware network function placement for efficient traffic handling in carrier cloud," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 2402–2407.

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

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

[20] 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.

[21] M. R. Sierra and C. A. C. Coello, *Improving PSO-Based Multi-Objective Optimization Using Crowding, Mutation and $\epsilon$-Dominance*. Berlin, Germany: Springer, 2005.

[22] E. Rodriguez, G. P. Alkmim, N. L. S. da Fonseca, and D. M. Batista, "Energy-aware mapping and live migration of virtual networks," *IEEE Syst. J.*, vol. 11, no. 2, pp. 637–648, Jun. 2017.

[23] G. P. Alkmim, D. M. Batista, and N. L. S. da Fonseca, "Mapping virtual networks onto substrate networks," *J. Internet Services Appl.*, vol. 4, no. 1, p. 3, Dec. 2013.

[24] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1888–1906, 4th Quart., 2013.

[25] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE mobile core gateways, the functions placement problem," in *Proc. 4th Workshop All Things Cellular, Oper., Appl., Challenges*, 2014, pp. 33–38.

[26] H. Ko, G. Lee, I. Jang, and S. Pack, "Optimal middlebox function placement in virtualized evolved packet core systems," in *Proc. 17th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Aug. 2015, pp. 511–514.

[27] 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./Oct. 2015, pp. 97–102.

[28] 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.

[29] R. Mijumbi *et al.*, "Design and evaluation of algorithms for mapping and scheduling of virtual network functions," in *Proc. IEEE Conf. Netw. Softwarization*, Apr. 2015, pp. 1–9.

[30] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, May 2015, pp. 98–106.

[31] A. Ksentini, M. Bagaa, T. Taleb, and I. Balasingham, "On using bargaining game for optimal placement of SDN controllers," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[32] 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.

[33] 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.

[34] 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.

[35] T. Taleb, A. Ksentini, and B. Sericola, "On service resilience in cloud-native 5G mobile systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 483–496, Mar. 2016.

[36] 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./Dec. 2017.

[37] 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./Nov. 2016, pp. 19–27.

[38] M. M. A. Khan, N. Shahriar, R. Ahmed, and R. Boutaba, "Multi-path link embedding for survivability in virtual networks," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 2, pp. 253–266, Jun. 2016.

[39] 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.

[40] R. Guerzoni, Z. Despotovic, R. Trivisonno, and I. Vaishnavi, "Modeling reliability requirements in coordinated node and link mapping," in *Proc. IEEE 33rd Int. Symp. Reliable Distrib. Syst.*, Oct. 2014, pp. 321–330.

[41] 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 Manag.*, vol. 14, no. 3, pp. 554–568, Sep. 2017.

[42] S. Ahuja and M. Krunz, "Algorithms for server placement in multiple-description-based media streaming," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1382–1392, Nov. 2008.

[43] M. Mangili, F. Martignon, and A. Capone, "Stochastic planning for content delivery: Unveiling the benefits of network functions virtualization," in *Proc. IEEE 22nd Int. Conf. Netw. Protocols*, Oct. 2014, pp. 344–349.

[44] 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.

[45] T. Furuta, M. Sasaki, F. Ishizaki, A. Suzuki, and H. Miyazawa, "A new clustering algorithm using facility location theory for wireless sensor networks," Nanzan Academic Soc. Math. Sci. Inf. Eng., Nagoya, Japan, Tech. Rep. NANZAN-TR-2006-04, 2007.

[46] N. Laoutaris, G. Smaragdakis, K. Oikonomou, I. Stavrakakis, and A. Bestavros, "Distributed placement of service facilities in large-scale networks," in *Proc. 26th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, May 2007, pp. 2144–2152.

[47] *Multimedia Broadcast/Multicast Service (MBMS); Architecture and Functional Description*, document TS 23.246, 3GPP, Jun. 2010.

[48] 5GXCast. *LTE-Advanced Pro Broadcast Radio Access Network Benchmark*. Accessed: Jan. 2018. [Online]. Available: http://5g-xcast.eu/wp-content/uploads/2017/12/5G-XcastD3.1v1.2web.pdf

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

[50] 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.

[51] T. Subramanya, L. Goratti, S. N. Khan, E. Kafetzakis, I. Giannoulakis, and R. Riggio, "A practical architecture for mobile edge computing," in *Proc. IEEE Conf. Netw. Function Virtualization Softw. Defined Netw. (NFV-SDN)*, Nov. 2017, pp. 1–4.

[52] M. J. Alves, "Using MOPSO to solve multiobjective bilevel linear problems," in *Proc. 8th Int. Conf. Swarm Intell.*, 2012, pp. 332–339.

[53] C. R. Raquel and P. C. Naval, Jr., "An effective use of crowding distance in multiobjective particle swarm optimization," in *Proc. 7th Annu. Conf. Genetic Evol. Comput.*, 2005, pp. 257–264.

[54] Y. Feng, B. Zheng, and Z. Li, "Exploratory study of sorting particle swarm optimizer for multiobjective design optimization," *Math. Comput. Model.*, vol. 52, nos. 11–12, pp. 1966–1975, 2010.

[55] A. J. Nebro, J. J. Durillo, J. Garcia-Nieto, C. A. C. Coello, F. Luna, and E. Alba, "SMPSO: A new PSO-based metaheuristic for multi-objective optimization," in *Proc. IEEE Symp. Comput. Intell. Multi-Criteria Decis.-Making (MCDM)*, Mar./Apr. 2009, pp. 66–73.

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

[57] X. Ge, S. Tu, G. Mao, and C. X. Wang, "5G ultra-dense cellular networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[58] Qualcomm. (Jan. 2017). *Augmented and Virtual Reality: The First Wave of 5G Killer Apps*. [Online]. Available: https://www.qualcomm.com/documents/augmented-and-virtual-reality-first-wave-5g-killer-apps

[59] P. Muñoz, O. Sallent, and J. Pérez-Romero, "Self-dimensioning and planning of small cell capacity in multitenant 5G networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4552–4564, May 2018.

[60] Y. Kim, J. Kwak, and S. Chong, "Dual-side optimization for cost-delay tradeoff in mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1765–1781, Feb. 2018.

**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 currently 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** 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 authored or co-authored over 400 papers and has supervised over 60 graduate students. He is the Senior Editor of *IEEE Communications Magazine*, and an Editorial Board Member of *Computer Networks* and *Peer-to-Peer Networking and Applications*. 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 a Vice President of Technical and Educational Activities of the IEEE ComSoc. He served as a ComSoc Vice President for Publications, a Vice President for Member Relations, the Director for Conference Development, the Director for Latin America Region, and the Director for On-Line Services. He is the Past Editor-in-Chief of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.