IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 13, 2020, accepted March 27, 2020, date of publication April 8, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986535

Optimal Assignment Plan in Sliced Backhaul Networks

CHRISTIAN QUADRI, (Member, IEEE), MARCO PREMOLI, ALBERTO CESELLI, SABRINA GAITO, (Member, IEEE), AND GIAN PAOLO ROSSI, (Member, IEEE)
Computer Science Department, University of Milan, 20133 Milano, Italy

Corresponding author: Christian Quadri (christian.quadri@unimi.it)

This work was supported by the Italian Ministry for Instruction, University and Research under the PON VASARI Project under Grant ARS01_00456.

ABSTRACT The 5G mobile network will rely on network slicing to provide a wide variety of services with various quality of service (QoS) requirements. Network slicing is promoted by 3GPP and provides a logical vertical partition of the network that is based on network virtualization technologies, namely, network function virtualization (NFV), software-defined networking (SDN) and ETSI multi-access edge computing (MEC). Despite the undisputed benefits in terms of flexibility and scalability that are pledged by the paradigm, network slicing requires intelligent resource scheduling and allocation algorithms to efficiently use the network resources, especially at the edge of the network, due to their scarcity. In this paper, we propose an optimization algorithm for steering data traffic of multiple slices in the edge backhaul network, which aims at maximizing the QoS. We extensively analyze the realizable grade of QoS by testing various levels of MEC resources, demonstrate the beneficial impact of the approach for mobile operators, and highlight the performance advantage that is realized versus a single-slice approach of undifferentiated traffic.

INDEX TERMS Multi-access edge computing, network slices, mathematical optimization.

I. INTRODUCTION

The next generation of mobile networks (5G) will support a wide variety of vertical services, each with specified quality of service (QoS) parameters. They will range from typical end-user services, such as video streaming and augmented and virtual reality (AR/VR), to Internet of things (IoT) applications, e.g., Industry 4.0 and smart cities. To handle such complexity and variety, Third-Generation Partnership Project (3GPP) has introduced the concept of network slicing, which refers to the creation of dynamic, logical and vertical partitions of the network to satisfy the requirements of specified service categories. Their implementation relies on the advances of network virtualization technologies, namely, network function virtualization (NFV) and software-defined networking (SDN), [1]. At the same time, the European Telecommunications Standards Institute (ETSI) has proposed the multi-access edge computing (MEC) [2], which offers cloud-computing capabilities at the edge of the network with the objective of reducing the network latency between end-users and the service.

The progressive softwarization of the network has led to the development of tools and platforms, such as management and orchestration (MANO), for managing the lifecycles of the slices, together with the underlying virtual network functions (VNFs) at the network levels; for example, ETSI standardizes the VNF architecture [3] and proposes the OpenSource MANO (OSM) [4] platform. The availability of these platforms significantly simplifies the sharing of resources among slices, but it still calls for the design of intelligent resource scheduling and allocation algorithms to enable a specified slice to satisfy its service level agreement (SLA) [5]. This problem is emphasized at the network edge, where the available resources are limited and the spatiotemporal dynamics of the traffic demand are high. According to the network edge structure that is illustrated in Fig. 1, specified actions must be taken in the radio access network (RAN) (see [6] for a comprehensive survey of the solutions for managing and orchestrating network slices in the access network) and in the backhaul network, where the traffic must be steered from/to the MEC layer without exhausting the limited MEC resources.

This paper addresses the problem of providing an optimized plan for managing the multi-slice traffic demand in

The associate editor coordinating the review of this manuscript and approving it for publication was Miguel López-Benítez.

the backhaul network. The problem combines two critical issues: the analysis of the spatiotemporal pattern of the traffic demand with various QoS levels and the exploitation of the results to proactively plan the network configuration to realize the optimal utilization of resources. The first issue has been widely studied in the literature [7], whereas the second remains an open problem for research. In the literature, most of the contributions [8]–[14] address this problem by attempting to minimize the network resources without violating the SLA. This paper address the problem via another strategy: maximizing the QoS of each slice without violating the available capacity. This approach has the advantages of providing the mobile operator with valuable indicators of the realizable performance for the specified level of capital expenditure (CAPEX)/operational expenditure (OPEX) and of issuing preemptive notifications that enable the prevention of the rise of spatiotemporal criticalities across the infrastructure.

The main contributions of the paper are threefold:

i) We establish a combinatorial optimization model that natively supports multiple network slices, which differ in terms of QoS requirements. The model starts from the single-slice model that is presented in [15] and expands it to incorporate the multi-slice characteristic of modern mobile networks. The algorithm addresses a combinatorial problem that is a multi-period variant of the generalized assignment problem.

ii) We extensively analyze the assignment plans by measuring the QoS level that is realizable by both single- and multi-slice optimization algorithms and by considering various levels of available resources at the MEC layer. We show that the network performance benefits from a multi-sliced approach that is more suitable for capturing the distinct spatiotemporal pattern of each slice than the previous single-slice model.

iii) We provide the mobile operators with a methodological framework for evaluating both the quality and resilience of their network infrastructure from the intensity of the CAPEX/OPEX investments. The proposed framework leverages the aggregated traffic only, thereby preserving the privacy of each subscriber.

To evaluate the performance of the multi-slice model, we consider a scenario with two network slices: one that has strict delay requirements, e.g., conversational voice, and another that lacks specified delay requirements, e.g., TCP-based traffic and text messages. To model the traffic demands of both slices, we exploit an anonymized mobile phone dataset that gathers the phone activities of approximately one million subscribers. The results demonstrate that the proposed model significantly improves the QoS and that the obtained plans can be exploited by the network operator both at a tactical level, to obtain valuable information about the effective dimensioning of the facility's capacity to realize a target level of QoS, and at an operational level, to associate base stations to facilities in new and unplanned network settings.

The remainder of this paper is organized as follows. Section II provides the background of this study. The multi-slice optimization model is presented in Section III, while in Section IV. we describe the simulation scenario. In Sections V and VI, we present the numerical results on QoS. Section VII presents the conclusions of the paper.

ACRONYMS

3GPP	Third-Generation Partnership Project
AR	augmented reality
BE	best- effort
CAPEX	capital expenditure
ETSI	European Telecommunications Standards Institute
IoT	Internet of things
MANO	management and orchestration
MCC	mobile cloud computing
MEC	multi-access edge computing
NFV	network function virtualization
OPEX	operational expenditure
OTT	over the top
QoS	quality of service
RAN	radio access network
RT	real time
SDN	software-defined networking
SFC	service function chain
SLA	service level agreement
V2X	vehicle to everything
vAP	virtual access point
vBS	virtual base station
VNF	virtual network function
VNF-RA	VNF resource allocation
VR	virtual reality

II. BACKGROUND

A. NETWORK SLICING

As a key feature of 5G networks, network slicing enables the creation of logical vertical partitions of the network to satisfy the requirements of specified service categories or even of a single vertical service [16]. Each slice is composed of a set of interconnected virtual network functions (VNFs), namely, the service function chain (SFC) [17], that implement the vertical service. Each component of the slice, namely, each node or link, has its own specific requirements, e.g., computational/storage capacity for nodes and delay/capacity for links, which must be satisfied for effective service provisioning. Fig. 1 illustrates the conceptual architecture of the network slices. In this example, we consider two autonomous network slices, namely, *Slice A* and *Slice B*, each of which has specified SFC and SLA requirements. As illustrated in the figure, the slicing process affects the entire mobile network infrastructure, from the core network down to the radio access network (RAN). The mobile operator embeds the slices' SFC into the physical infrastructure and guarantees both data plane isolation and the satisfaction of the SLA requirements [5].

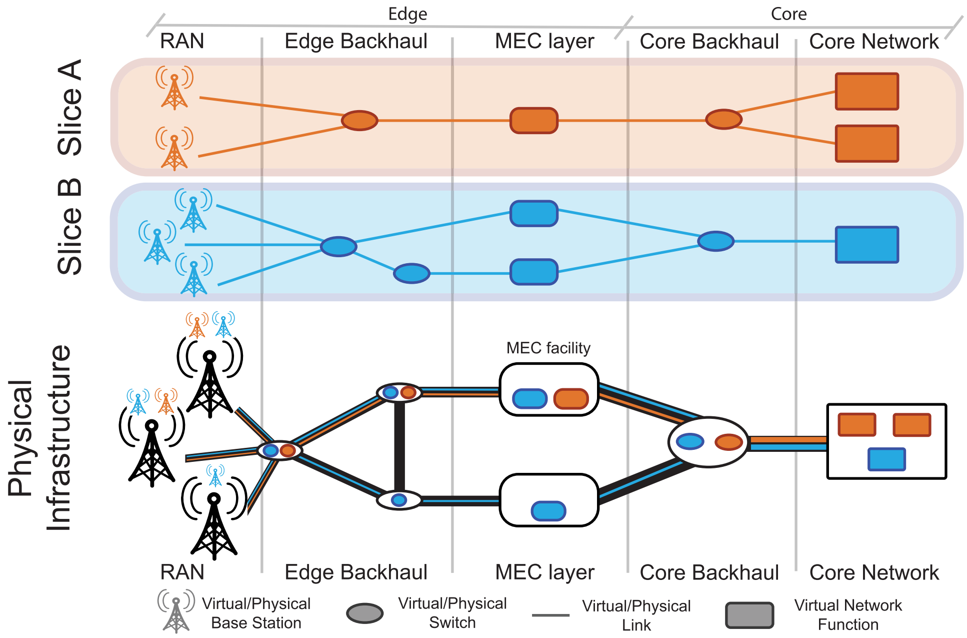


FIGURE 1. **Conceptual architecture of the network slices.**

The dynamics of the traffic that is associated with each slice requires elastic resource allocation to either reserve more resources or release them according to the traffic load. As shown in [18], network slicing has deep implications with respect to resource management; hence, network operators are seeking a trade-off between offering a fully customized network partition (that satisfies the service requirements perfectly) and the efficient allocation of its own available resources. The problem to be solved at this level is to ensure elastic resource allocation by satisfying the specified SLA while optimizing the placement of the VNFs within the variety of slices. Various approaches for addressing VNF resource allocation (VNF-RA) are presented in [19]. The optimization can be conducted by using mixed-integer linear programming models [8], [9], [15], heuristics algorithms [10], [11], game theory [12], [13] or machine learning [14] approaches.

B. MULTI-ACCESS EDGE COMPUTING

Cloud computing has been offering a successful computational model for many years because it ensures a large amount of resources, high availability and service elasticity through virtualization. Nevertheless, cloud computing is becoming increasingly unsuitable for many emerging applications, such as AR/VR, massive IoT deployment and ultra-reliable communications. These new applications, which utilize the 5G technology, all rely on various latency and reliability constraints and, as in case of IoT, generate a huge amount of uplink traffic that ultimately floods the backhaul network. To support this new class of application scenarios, ETSI proposed multi-access edge computing (MEC) [2], which offers cloud-computing capabilities that are distributed at the edge of the network. In addition to reducing the communication delay, MEC is endowed with peculiar features such as proximity- and context-awareness and geo-localization, which are difficult to realize in a traditional cloud environment. However, a few shortcomings offset these advantages. They include, for instance, far lower availability of computational and storage capabilities at each MEC node compared to those that are offered by any cloud platform. This limitation motivates the design of a radically new resource management strategy because the resource orchestration policies that are commonly adopted by cloud operators are unsuitable in an MEC scenario, where the resources are highly distributed and the traffic load at the edge is highly dynamic and non-homogeneous.

Despite the recent introduction of MEC, many contributions have been produced by the research community over the last few years. In the early literature, MEC was regarded as an extension of mobile cloud computing (MCC) that provides offloading capabilities at the edge of the network. [20] presents a comprehensive survey of such use cases of MEC. When MEC is employed to support IoT and smart city scenarios, the literature on MEC overlaps with the contributions on FOG computing, according to [21], where a complete overview of the MEC/FOG literature is provided. The convergence between 3GPP and ETSI-MEC has led to the inclusion of MEC into the ecosystem of 5G as a promising solution for bringing computation capacities to the edge [22].

C. NETWORK SLICES AT THE EDGE

The management of the network slicing at the core layer benefits from a large amount of computational resources, accurate prediction of the traffic dynamics, and a highly

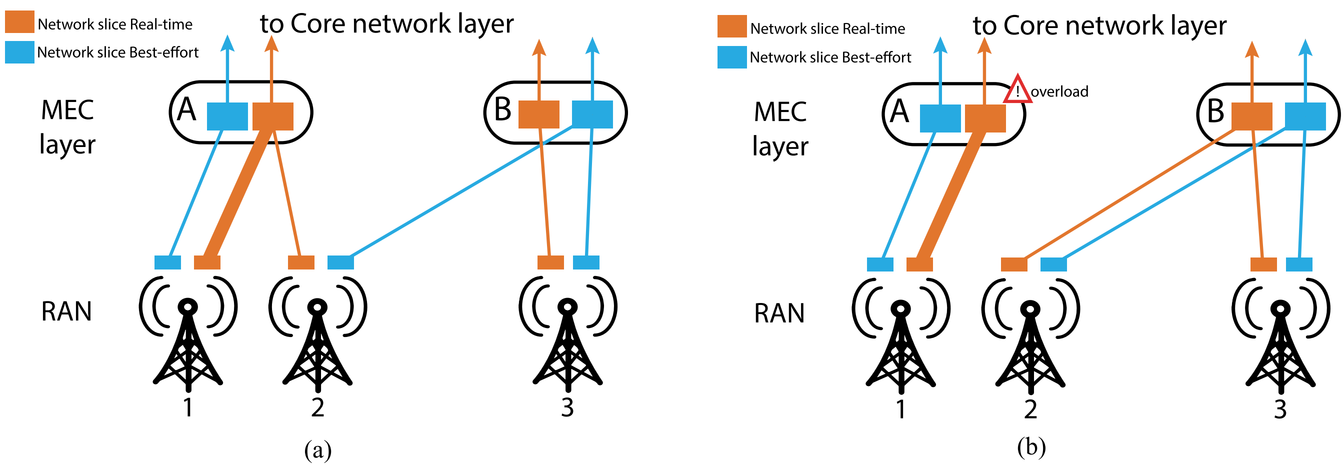


FIGURE 2. Example of base station-facility association: (a) The real-time slice is always associated with the closest facility, and (b) the facility overload condition causes part of traffic to be associated with another facility.

centralized topology (few datacenters cover a vast geographical area) and references the well-established literature on the optimization of the placement of VNFs [23]. However, at the edge of the network, such management becomes challenging. As shown in [18], the sharing efficiency decreases at the edge of the network. As shown in Fig. 1, the network edge is composed of three sub-layers: a RAN, an edge backhaul and an MEC. At the RAN level, spectral, processing and scheduling resources are shared among multiple autonomous slices that differ in terms of QoS requirements and traffic load dynamics. Reference [6] provides a comprehensive survey of the challenges in and approaches for managing and orchestrating network slices in the access network. The edge backhaul steers the downlink/uplink traffic to/from the RAN and must satisfy the two following constraints: (i) all downlink/uplink data traffic must be handled by the backhaul network, namely, offloading onto other networks is not allowed; and (ii) the amount of the traffic load that is steered to/from an MEC facility should not exceed its limited amount of available resources. In Fig. 2a, we present the case of two network slices that share resources at the edge of the network and support different classes of applications. The blue slice handles a bundle of best-effort data flows, e.g., buffered streaming and web browsing, whereas the orange slice manages real-time data traffic, e.g., video conferencing, online gaming, and V2X communications. The distance between the RAN and the serving-edge facility must be considered if the low latency requirements of the orange slice must be satisfied. Thus, the QoS requirements of the real-time slice would more likely be satisfied if the relevant traffic to/from the RAN were managed by the closest MEC facility, while the blue slice is almost independent of similar distance constraints.

Nonetheless, the limited resources of MEC facilities are a critical issue for the resource orchestration process. Unfortunately, the straightforward assignment of real-time traffic to the closest facility could lead to an exhaustion of the resources at the facility, thereby hindering the realization of the required SLA. Such a condition is described in Fig. 2b, where the overload condition of facility A causes the diversion of part of the real-time traffic - to/from base station 2 - from facility A to facility B. This leads to a new design problem: the identification of an assignment that satisfies the SLA constraints without exceeding the facility's capacity.

The variation in the traffic demand over time may render appealing (to provide better QoS), or even necessary (to avoid the exhaustion of resources at the facility), the modification of the base station-facility assignments. Reference [24] relies mainly on the signalling traffic needed to reconfigure the network path, the migration of state-full VNFs, and the reservation of resources at the target facility. Therefore, the mobile operator must identify the optimal trade-off between the pursuit of QoS optimality, which may cause frequent migrations, and the minimization of the number of switches, which leads to sub-optimal assignments.

III. MULTI-SLICE EDGE ORCHESTRATOR

In this study, we assume a sliced network scenario, as illustrated in Fig. 1, and we focus on the optimization of QoS at the edge backhaul level by managing and orchestrating the virtual links between RAN and MEC VNFs of the slice. We assume that the mobile operator is entirely responsible for slice management and the provisioning of a set of slices that differ in terms of QoS parameters. Over the top OTT service providers are supposed to share slices according to the service requirements. In such a scenario, each slice aggregates different traffic flows that are rooted on a variety of OTT services with similar QoS requirements and negotiated SLAs.

A. REAL-TIME AND BEST-EFFORT SLICES

We consider two network slices, namely, *real time (RT)* and *best- effort (BE)*, which are designed to support opposite service types. The RT slice is assumed to satisfy strict latency constraints and to ensure a short response time. In [25], 3GPP clearly defines the packet delay budget for a variety of applications with real-time constraints of variable strictness. For instance, the delay budget is set to ≤ 50 ms for hard real-time traffic (AR/VR, V2X, or live interactive gaming) and to 100 ms for conversational voice traffic. In contrast, the BE slice is assumed not to be constrained by strict requirements.

1) SLICES AT THE RAN

At the RAN level, we adopt the notion of virtual access point (vAP), which was introduced in [26], and we model each slice as a VNF-based system. A vAP is a RAN VNF, which manages the radio resources for a slice and is deployed on a physical access point. Each vAP is managed by the orchestrator and can be dynamically deployed/undeployed according to the dynamics of the slice traffic. Likewise, we model the RAN level of each slice as a set of virtual base station (vBS), namely, a VNF that is deployed on a physical base station and is responsible for handling the users' generated traffic load. The deployment of vBSs relies on the Cloud-RAN architecture [27] and virtualization technology [28] for the allocation of the necessary amount of resources (spectral, processing and scheduling resources) for satisfying the QoS requirements. We assume that an instance of vBS of a network slice is deployed on a physical base station if at least one user, who is connected to the physical base station, uses the slice. Moreover, we assume that the amount of resources that are allocated for managing the slice is directly proportional to the amount of traffic load that is generated by users who are currently using the slice.

2) SLICES AT THE EDGE BACKHAUL

According to Fig. 1, the physical edge backhaul is composed of a set of interconnected SDN switches, which can deploy a virtual path throughout a set of OpenFlow data plane rules, namely, the mobile operator can map virtual links onto a physical forwarding path to/from MEC facilities. To exploit the distributed nature of the MEC layer and the horizontal scaling of the VNFs, namely, the deployment of multiple replicas of the same VNF on different MEC facilities, we assume that a single virtual link of the slice SFC can be mapped to multiple physical forwarding paths. As a consequence, the uplink traffic that is issued by a vBS can be forwarded to various MEC facilities, while the downlink traffic can arrive from multiple MEC VNFs that are deployed in various facilities.

3) SLICES AT MEC

In this study, we rely on the MEC paradigm that was proposed by ETSI, which offers cloud-computing capabilities at the edge of the network, thereby reducing the end-to-end latency [2]. Some applications require special hardware for proper operation, e.g., AR/VR applications utilize GPU. To support this class of applications, MEC facilities should be equipped with specialized hardware, the cost of which could exceed the cost of general-purpose hardware. Consequently, the network operator may decide to equip only a subset of servers inside a facility with special hardware capabilities. We model such a feature by assuming that only a fraction of the capacity in each facility can be used by the RT slice. These resources are not reserved for RT traffic only, but priority is given to it: resources can be allocated to BE traffic only if they are not being fully utilized by RT traffic. The portion of the capacity on which the RT traffic has a priority is a parameter of our model, and its value should depend on the ratio between RT and BE traffic, namely, the higher the ratio, the higher the value.

B. NOTATION AND MODELING

In the following, we fully formalize our modeling choices. This mathematical formalization step is inspired by [15], in which a *single-slice* MEC optimization problem is considered.

The main strategy behind our mathematical model is to map input data (vBS demands, facility capacities, and assignment and migration costs) to optimal decisions. We encode both the assignment of vBSs to facilities and the migration from facility to facility, which are decisions that an operator must repeatedly take over time, as output solution variables, both for BE and RT slices. The mapping is formalized by using mathematical optimization notation [29], including both a set of requirements that any solution must satisfy and an aspiration criterion. The requirements are as follows: no facility can manage a vBS load that exceeds its capacity, each vBS must be assigned to one facility, and assignment and migration decisions must be consistent with each other. The aspiration criterion minimizes a linear combination of the assignment and migration costs.

Formally, we assume that the planning time horizon has been discretized; in a practical scenario, the discretization may match the granularity of the input data. Let T be the set of time slots that arise in such a discretization. Let A^{BE} and A^{RT} be the sets of virtual base stations of the best-effort and real-time slices, respectively. We define $A = A^{BE} \cup A^{RT}$ as the set of all vBSs that are deployed on physical base stations. Let K be the set of MEC facilities.

We suppose that the following data are available:

- w_i^t: the demand of vBS $i \in A$ during time slot $t \in T$, which is the amount of traffic of users who connect to i
- C_k (resp. C_k^{RT}): the overall (respectively, RT) capacity of facility $k \in K$, namely, the maximum amount of overall (respectively, RT) traffic that can be serviced by k in each time slot
- m_{ik}: the assignment cost of vBS i to facility k
- l_{jk}: the migration cost from facility j to facility k.

Our goal is to identify effective resource allocation plans, which are formally defined as follows:

- decision variable $x_{ik}^t \in [0, 1]$ encodes assignment, which represents the fraction of traffic from vBS i that is assigned to facility k at time t

- decision variable $y_{ijk}^t \in [0, 1]$ encodes migration, which represents the fraction of the traffic from vBS i that must be switched from facility j to facility k at time t.

Table 1 summarizes the notation that is adopted in the optimization model.In a resource allocation plan, the BE slice structure is finally defined by considering only the x_{ik}^t and y_{ikl}^t variables for which $i \in A^{BE}$. Similarly, the RT slice structure is defined by the x_{ik}^t and y_{ikl}^t variables for which $i \in A^{RT}$.

TABLE 1. Notation table of the optimization model.

Notation	Description
T	set of time slots
A^{RT}	set of the vBSs of the real-time slice
A^{BE}	set of the vBSs of the best-effort slice
A	set of all vBSs, $A^{BE} \cup A^{RT}$
K	set of facilities
C_k	overall capacity of facility $k \in K$, namely, the maximum amount of overall traffic that can be serviced by k in each time slot
C_k^{RT}	maximum capacity of facility $k \in K$ that can be used by the real-time slice in each time slot, $C_k^{RT} \leq C_k$
w_i^t	demand of vBS $i \in A$ during time slot $t \in T$, which is the amount of traffic of users who are connecting to i
m_{ik}	assignment cost of vBS i to facility k
l_{jk}	migration cost from facility j to facility k
x_{ik}^t	decision variable $\in [0,1]$ that encodes assignment, which represents the fraction of traffic from vBS i that is assigned to facility k at time t
y_{ijk}^t	decision variable $\in [0,1]$ that encodes migration, which represent the fraction of the traffic from vBS i that must be switched from facility j to facility k at time t
α	parameter of the relative importance of migration (default 1)
β	parameter of the relative importance of assignment (default 1)

The identification of optimal resource allocation plans from data is not trivial: it requires the solution of an optimization problem that is even more general than that approached in [15]. The additional complexity arises from the necessity of considering two overlapping types of traffic, while the algorithms of [15] are suitable for a single traffic type only. However, we managed to extend the models of [15] and to re-design the optimization algorithms such that the additional problem complexity is addressed with a minimal increase in the algorithmic complexity. We formulate the problem of optimally assigning vBSs to facilities over time as follows:

$$\min \sum_{t\in T}\sum_{i\in A}\left(\alpha \sum_{\substack{(j,k)\in\\ K\times K}} w_i^t l_{jk} y_{ijk}^t + \beta \sum_{k\in K} w_i^t m_{ik} x_{ik}^t\right) \quad (1)$$

$$\text{s.t.} \sum_{i\in A} w_i^t x_{ik}^t \leq C_k \quad \forall t \in T,\ \forall k \in K \quad (2)$$

$$\sum_{i\in A^{RT}} w_i^t x_{ik}^t \leq C_k^{RT} \quad \forall t \in T,\ \forall k \in K \quad (3)$$

$$\sum_{k\in K} x_{ik}^t = 1 \quad \forall i \in A,\ \forall t \in T \quad (4)$$

$$x_{ik}^t = \sum_{l\in K} y_{ilk}^t \quad \substack{\forall i\in A, \forall k\in K\\ \forall t\in T\setminus\{1\}} \quad (5)$$

$$x_{ik}^t = \sum_{l\in K} y_{ikl}^{t+1} \quad \substack{\forall i\in A, \forall k\in K\\ \forall t\in T\setminus\{T\}} \quad (6)$$

$$x_{i,k}^t \in [0, 1],\ y_{i,k,k'}^t \in [0, 1] \quad \substack{\forall i\in A, \forall t\in T\\ \forall k,k'\in K} \quad (7)$$

Formally, a logical connection is required at time t between each vBS i and each facility k such that $x_{ik}^t > 0$, whereas the actual value of x_{ik}^t represents the fraction of traffic to be sent from i to k. Variables $y_{i,k,k'}^t$ have a similar interpretation. The objective function (1) contains the sum of two terms, which model the migration and the assignment costs. Parameters α and β are assumed to be constants, which must be set by the network planner to fine tune the relative importance of assignment and migration in the final QoS (we refer to Section IV for a general discussion on suitable parameter settings for simulations). Conditions (7) define the domain of each decision variable. Constraints (2) and (3) ensure that the capacity of a facility is never exceeded: (3) ensures that the amount of RT traffic that is assigned to the facility does not exceed its RT capacity. In contrast, constraints (2) consider both RT and BE traffic, namely, BE traffic can use the residual resources of k up to its overall capacity. Constraints (4) have two roles: First, together with non-negativity conditions on x_{ik}^t, they ensure that every vBS is logically connected to at least one facility in every time slot; second, they ensure that in each time slot, all the traffic for each vBS is assigned to facilities, potentially by splitting. Constraints (5) and (6) ensure that the assignment and migration decisions are consistent.

The BE and RT decisions are defined by different sets of variables, which are linked by capacity constraints (2).

In Fig. 3, we illustrated how the mathematical formulation of the assignment problem leads to the orchestration of multiple slices. We consider the case of a single time slot and two vBSs that are deployed on a physical base station. Each vBS is associated with a virtual switch (vSwitch) VNF, which handles the traffic demand of the vBS and steers it toward distinct facilities according to the values of the assignment decision variables (x_{jk}^t and x_{im}^t in the figure). An analogous schema is defined for the migration among MEC facilities, but it is based on the values of the migration decision variables.

C. OPTIMIZATION ALGORITHMS

From a computational complexity perspective, the model is a linear program (LP); therefore, according to classical linear programming theory results, it is solvable in polynomial time (we refer the reader to [29] and [30] for all formal results on the subject). From a practical resolution perspective, however, its size renders it unmanageable for direct optimization algorithms. Indeed, in a preliminary round of experiments, we attempted to use state-of-the-art solvers such as [31], which halted already in a preprocessing phase. This is in line with computational experience that was reported in previous studies, such as [15].

In a few cases, models from the literature encode combinatorial structures. For instance, the model of [15] is a minimum

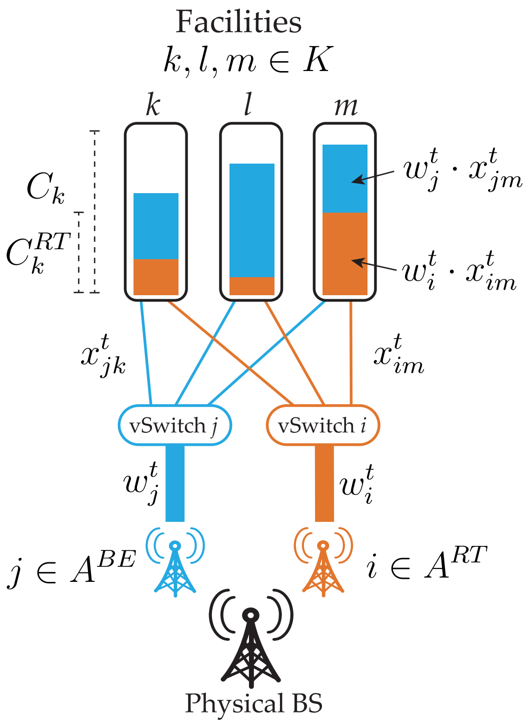


FIGURE 3. **Multi-slice network model of a single physical BS in a specified time slot *t*.**

cost flow problem, to which highly efficient algorithms can be applied.

Unfortunately, this is not the case when multiple slices are considered: constraints (2) and (3) render our structure more complicated. Nevertheless, we managed to devise an ad hoc algorithm by exploiting the so-called Dantzig-Wolfe decomposition principle [32]. A similar approach has been used in [15]. To keep the paper compact but self-contained, in the following, we present only the essential technical details that are novel and specific to our approach, and we refer to [32] and [15] for general descriptions of the theoretical and algorithmic frameworks that we employ.

Let S^i be the set of all possible sequences of assignments and migrations of vBS i to facilities over time:

$$S^i = \{(x_{ik}^t, y_{ilk}^t) : (4), (5), (6), (7)\}.$$

According to linear programming theory, such a set geometrically corresponds to a polyhedron: each point in S^i can be represented as a linear convex combination of the extreme points (and rays) of S^i. Let us denote as $(\bar{x}_{tk}^s, \bar{y}_{tlk}^s)$ these extreme points and as σ^i the corresponding set. For each $i \in A$:

$$(x_{ik}^t, y_{ilk}^t) = \sum_{s \in \sigma^i} (\bar{x}_{tk}^s, \bar{y}_{tlk}^s) \cdot z^s, \ \sum_{s \in \sigma^i} z^s = 1$$

where z^s are new decision variables, which represent the multipliers in the linear combination. For each extreme point $s \in \sigma^i$, let:

$$c^s = \sum_{t \in T} \left(\alpha \sum_{\substack{(j,k) \in \\ K \times K}} w_i^t l_{jk} y_{ijk}^t + \beta \sum_{k \in K} w_i^t m_{ik} x_{ik}^t \right)$$

After substitution according to these equations, model (1)–(7) becomes:

$$\min \sum_{i \in A} \sum_{s \in \sigma^i} c^s z^s \tag{8}$$

$$\text{s.t.} \sum_{i \in A} \sum_{s \in \sigma^i} w_i^t \bar{x}_{tk}^s z^s \leq C_k \quad \forall t \in T, \ \forall k \in K \tag{9}$$

$$\sum_{i \in A^{RT}} \sum_{s \in \sigma^i} w_i^t \bar{x}_{tk}^s z^s \leq C_k^{RT} \quad \forall t \in T, \ \forall k \in K \tag{10}$$

$$\sum_{s \in \sigma^i} z^s = 1 \quad \forall i \in A \ \forall i \in A \tag{11}$$

Model (8)–(11) is still a linear program, but now it contains one variable for each element of σ^i, and these variables grow combinatorially with respect to $|K|$ and $|T|$. Although each σ^i encodes sequences of assignments and migrations and, therefore, a special shortest path structure, there is no guarantee that a globally feasible solution to (8)–(11) can be obtained by independently solving a shortest path problem for each vBS i in A as the sequences interact with one another due to constraints (9) and (10).

However, its optimization can be pursued iteratively via dynamic variable generation techniques [32], [33]: we replace each σ^i with an arbitrarily small $\bar{\sigma}^i \subset \sigma^i$, we solve the restricted problem that is obtained via this approach, we collect the corresponding dual solution and we determine whether the variables with negative reduced cost exist, which correspond to elements in $\sigma^i \setminus \bar{\sigma}^i$. If they do not, then the solution of the restricted problem is optimal *for the full problem* as well; otherwise, a few of these potentially useful variables are added to the sets $\bar{\sigma}^i$, which grow incrementally, and the whole process is iterated.

The search for new variables is conducted implicitly by solving at each iteration an *optimization* (pricing) subproblem and searching for the variable that has the most negative reduced cost. In our case, such a search is conducted by solving one shortest path problem for each $i \in A$ in a directed acyclic graph that has one layer for each $t \in T$ that contains one vertex for each $k \in K$. Arcs connect each vertex in layer t to all vertices in layer $t + 1$. Its structure is illustrated in Figure 4. Each arc from a vertex $(j, t-1)$ to a vertex (k, t) has an associated weight $\bar{w}_{j,t-1,k,t}$, which accounts for (a) the cost of assigning i to j at time $t-1$, (b) the cost of assigning i to k at time t, (c) the migration cost between j and k at time t, (d) the contribution of the dual variables that are associated with constraints (9) and, only if $i \in A^{RT}$, (e) the contribution of the dual variables that are associated with constraints (10). Each shortest path problem is solved highly efficiently by a dedicated dynamic programming algorithm, as outlined in Algorithm 1. For each $i \in A$, when dynamic programming has been completed, we consider the minimum $\bar{c}_{k,|T|}$ value

Algorithm 1: Exact Pricing Algorithm

```
Result: Optimal planning over time for vBS $i$
init: for $k \in K$ do
    $\bar{c}_{k,0} \leftarrow 0$;
end
solve: for $t \leftarrow 1$ to $|T|$ do
    for $k \in K$ do
        $\bar{c}_{k,t} \leftarrow \min_{j \in K}\{\bar{c}_{j,t-1} + \bar{w}_{j,t-1,k,t}\}$;
        $\mathrm{pred}_{k,t} \leftarrow \mathrm{argmin}_{j \in K}\{\bar{c}_{j,t-1} + \bar{w}_{j,t-1,k,t}\}$;
    end
end
```

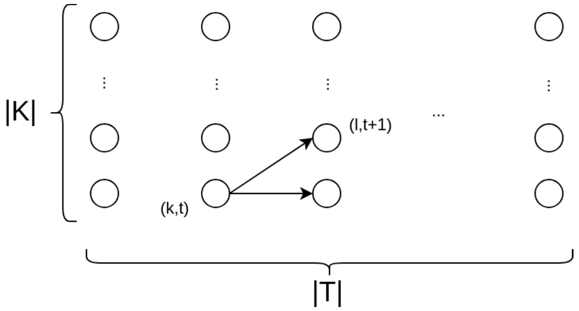


FIGURE 4. Structure of the pricing subproblem graph.

and rebuild a corresponding path by recursively following $\mathrm{pred}_{k,t}$: it encodes the column of the most negative reduced cost. In summary, our choice of formulation for the set of vBSs and constraints (2) and (3) enables the complexity of each shortest path subproblem to be maintained while simply requiring the solution of a potentially larger number of the subproblems (possibly even in parallel, as they are disjoint).

Additionally, from a computational complexity perspective, even if $|\sigma^i|$ is not polynomially bounded, such a procedure is still of polynomial time complexity. According to classical linear programming duality results, generating new variables in (8)–(11) is equivalent to finding violated cuts in its dual; the separation problem in the dual is equivalent to the pricing problem in the primal, which, in our case, can be solved in polynomial time. The equivalence between separation and optimization [29] implies that the full problem (8)–(11) can be solved in polynomial time.

We also report that, experimentally, we identified that such a procedure requires far fewer computing resources than the direct optimization of (1)–(7), in terms of both memory and CPU time. This accords with previous results from the literature [15].

Finally, as discussed in Section VI, our extension trades a minor increase in the algorithmic complexity for a major increase in the solution quality.

D. QOS MODELING

In line with [15], the methodology that we propose involves (a) population of the described model with estimates of the traffic demand and capacities from historical data and (b) optimization over a specified restricted time horizon that represents a pattern of periodicity (e.g., over one week). Therefore, the solutions that the model provides are *patterns*, which are assumed to be subsequently and periodically applied to new and unplanned network settings.

The quality of the pattern that is obtained by applying the optimization algorithm still provides only an indirect measure of the QoS that is offered by the network operator. We assume that the performance a service can realize by using a network slice is affected by two main factors: the load of the facility and the network delay that is associated with the physical path between the vBSs and the MEC VNFs. The first factor accounts for the computational costs at the MEC level, whereas the latter mainly depends on the path lengths between the base stations and MEC facilities. Due to the topology of the mobile operator backhaul network [34], the longer the path, the more network devices (i.e., routers and switches) to pass through, and the longer the related delay. Thus, in the following, we focus on the delay between physical base stations and MEC facilities as the main performance indicator of the MEC approach. This argument is emphasized when addressing the real-time slice.

Formally, we define the QoS factor that is related to the loads of the facilities (*QoS-Load*) as follows:

$$\text{QoS-Load}_i^t = \sum_{k \in K_i^t \subseteq K} \min\{e^{C_k - w_k^t}, 1\} \cdot x_{i,k}^t \tag{12}$$

where $K_i^t \subseteq K$ is the set of facilities to which the vBS is assigned, w_k^t is the actual load of facility k at time slot t, and C_k is capacity of facility k. According to this definition, when the actual load reaches the warning threshold, performances drop because the MEC facility becomes unable to provide service of suitable quality. For example, users can experience a long latency time due to congestion at the facility level. Regarding the network delay, we define the network QoS (*QoS-Net*) that is associated with vBS i at time t as follows:

$$\text{QoS-Net}_i^t = \sum_{k \in K_i^t \subseteq K} e^{-\frac{m_{i,k} - m_{i,k_i^*}}{\delta}} \cdot x_{i,k}^t \tag{13}$$

$$k_i^* = \operatorname*{argmin}_k m_{i,k} \tag{14}$$

where k_i^* is the best facility for vBS i, namely, the facility that is associated with the minimum latency, and δ is a scale parameter. According to the definition, QoS-Net is optimal, namely, is equal to 1, if and only if all the traffic load of the vBS is assigned to the best facility.

According to the previous definitions, we defined the QoS that is realized by the network operator from vBS i at time t for the two network slices (RT and BE) as follows:

$$QoS_i^t = \begin{cases} \text{QoS-Load}_i^t \cdot \text{QoS-Net}_i^t & \text{if RT slice} \\ \text{QoS-Load}_i^t & \text{if BE slice} \end{cases} \tag{15}$$

For the BE slice, the QoS is only affected by the load of the facilities because the BE slice is not sensitive to the network delay. Finally, a scale factor is added to express QoS in terms of the mean opinion score (MOS), which ranges between 1 and 5.

IV. SIMULATION SCENARIO

A. DATASET

Our data source consists of call detail records (CDRs) that describe the phone activities of approximately one million subscribers to one of the largest Italian mobile operators in the metropolitan area of Milan (surface of 235 km^2 (15.9 km × 14.8 km)) for a time period of 67 days (9 weeks) [35]. The dataset contains approximately 107 million calls (VoIP) and 52 million text messages. Each record contains the anonymized identities of the customers who were involved in the phone activity, the timestamp, the base station identification number where the subscriber was registered when starting the activity and the location area to which the base station belongs. The location area consists of a label that specifies a place in the city, e.g., a street, square, or station. Our dataset contains 224 location areas, which group 1663 base stations. In the call records, the duration of each call in seconds is also specified; based on this attribute, we can determine whether or not a call record corresponds to a missed call. Our dataset contains approximately 41 million (39%) missed call records. In Table 2, we describe in detail each field of the records in the CDR dataset. Since the dataset does not contain the MOS of the phone activities from a base station, we use expression (15) to estimate the average QoS that is provided to the users.

TABLE 2. Description of the fields of the CDR dataset.

Field	Description
	Call record
`caller`	anonymized identity of the customer who performed the call
`callee`	anonymized identity of the customer who received the call
`timestamp`	date and hour when the call was performed
`duration`	duration of the call in seconds (0 seconds correspond to no answer)
`base station id`	base station identification number where the caller was registered when starting the call
`location area`	label that specifies the place in the city, e.g., street, square, or station, where the base station is deployed.
	Text message record
`sender`	anonymized identity of the customer who sent the text
`destination`	anonymized identity of the customer who received the text
`timestamp`	date and hour when the text was sent
`base station id`	identification number of the base station to which the sender subscriber was registered when the text was sent
`location area`	label that specifies the place in the city, e.g., street, square, or station, where the base station is deployed.

B. PHYSICAL NETWORK INFRASTRUCTURE

In this section, we define the network topology, and we estimate the base station locations, the structure of the aggregation rings and the locations of the MEC facilities.

1) BASE STATIONS POSITIONING

Due to the sensitivity of the information on the positions of the base stations, the mobile operator does not provide us with the exact GPS coordinates of the base stations (BSs). To estimate the GPS position of each BS, we exploit the location area information that is associated with each BS. Each area aggregates adjacent BSs into groups of size 4 to 25, where the larger the group, the smaller the coverage radius of the BSs that belong to it. By exploiting the Google geocoding service, we obtain the approximated GPS position of the center of each location area. Then, we conduct Voronoi tessellation to determine the portion of the city that is occupied by each location area. From the Voronoi area of a location area, we generate a uniform grid of points with a distance of 250 meters, and we cluster these points via the K-means clustering algorithm by setting the parameter k to the number of BSs that belong to that location area. Finally, we select the resulting centroid position as the GPS coordinates of the BSs that belong to that location area. In Fig. 5a, we report the results of this process, where the areas are determined via Voronoi tessellation, while the red dots are the computed BS GPS positions.

2) AGGREGATION RINGS AND FACILITY POSITIONING

In accordance with [34], we assume that the infrastructure of the backhaul network is organized into hierarchical rings and divided into three segments: (i) access, (ii) aggregation, and (iii) core. Each ring is associated with a gateway (or a set of gateways for redundancy) that interconnects the upper/lower layers of the hierarchy, and an MEC-point of presence (MEC-PoP) (which satisfies the strict latency requirements of tactile interactions in an urban environment) is suitably located within the aggregation segment of the backhaul network. In accordance with the delay budget that we assigned to the RT slice, we assume that the facilities are directly connected to an M2 MEC PoP [34].

Given the backhaul network topology, we must connect the base stations to the aggregation rings where the facilities are deployed without knowing the real backhaul network topology that is deployed by the mobile operator. To conduct the task, we construct a simplified backhaul network by considering the trade-off between the distances between the aggregation nodes (M2 [34]) and the number of base stations that are connected to the same M2 node. We assume that the backhaul network is composed of 20 aggregation rings and 20 facilities; in accordance with [34], [36], this value represents a suitable number that is based on the number of the BSs that are connected to an aggregation ring. A straightforward approach for organizing the BSs in rings is to use the K-means algorithm to cluster the BSs in 20 clusters. However, as shown in Fig. 5a, the distribution of the BSs is not homogeneous; thus, the direct use of the clustering algorithm leads to a biased distribution of the facilities that overloads the facilities in the city center. To overcome this problem, we do not conduct clustering on the original BS positions but

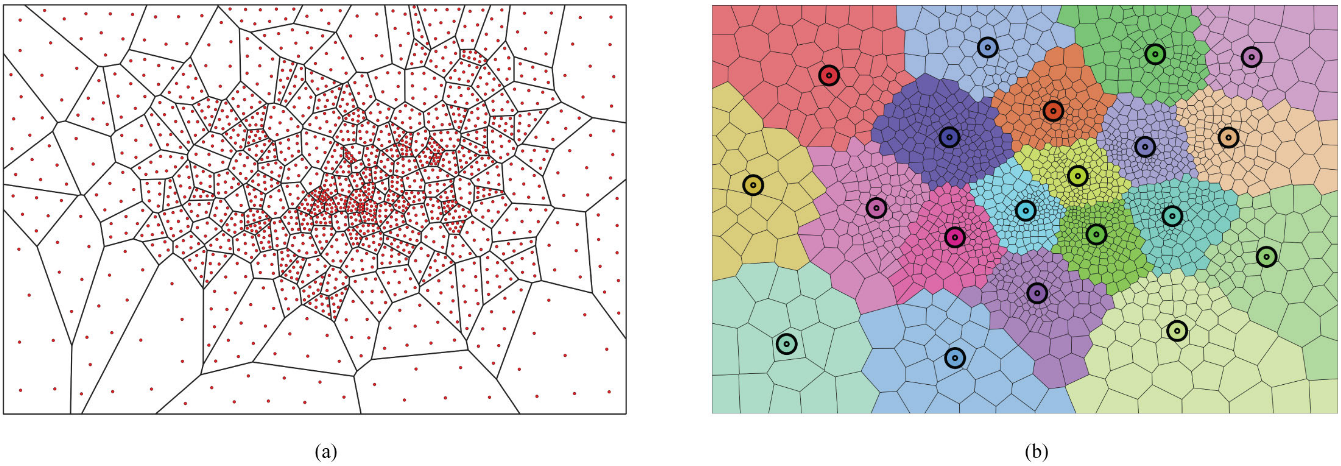


FIGURE 5. (a) A map of the Voronoi tessellation of 224 city areas and GPS base station positions (red dots). (b) A map of the positions of the 20 facilities, along with the areas of the clusters.

on a new set of coordinates that are obtained by applying an RBF-kernel.[1] The result is reported in Fig. 5b, where the circles represent the facilities (co-located with the M2 nodes), the BS areas are determined via Voronoi tessellation of the BS GPS positions, and the color of the area indicates the aggregation ring to which the BS is connected. The main implication of such a network topology is that the traffic to/from a vBS that is associated with a facility in the same aggregation ring (the centroid of the cluster in Fig. 5b) experiences the shortest delay; by contrast, if a non-optimal association is deployed, the traffic must be routed through the ring hierarchy, thereby resulting in longer delays and poor QoS-Net, as expressed by Eq. 13.

C. REAL-TIME AND BEST-EFFORT SLICE COST MODELING

1) PHYSICAL PATH COST

Both the assignment and migration costs depend directly on the network delay that derives from transmitting on the links between physical base stations and facilities and on the inter-facility communication path. These costs may increase stepwise according to the number of hops in the communication path. From a modeling perspective, our formulation can embed any cost function since these costs can be computed from the data during preprocessing and encoded as numerical coefficients in (1)–(7). From an experimental perspective, in the absence of data on the physical network topology, we make the probabilistic assumption that the *expected* number of hops and, therefore, the expected network delay, increases proportionally to the geographic distance among the communicating hosts (BSs and facilities). We assume that the longer the distance, the higher the expected number of network devices (i.e., routers and switches) to traverse, and the longer the expected network delay. Therefore, given the location of BSs and facilities, as in the previous section, we define $d_{i,f}$ as the distance between physical BS i and facility f of the aggregation ring to which the BS belongs. Similarly, we define $g_{f,h}$ as the distance between facilities k and l. In accordance with the topology of the backhaul network, we define the cost of the physical path that connects physical BS i to facility k, which is denoted as $p_{i,f}$, as follows:

$$p_{i,f} = d_{i,f_i^*} + g_{f_i^*,f} \qquad (16)$$

where f_i^* is the facility on the aggregation ring of BS i. If $f = f_i^*$, then $g_{f_i^*,f} = 0$; nonetheless, if the facilities differ, $g_{f_i^*,f}$ makes a smaller contribution to the definition of the value of $p_{i,f}$.

2) SLICE ASSIGNMENT COST

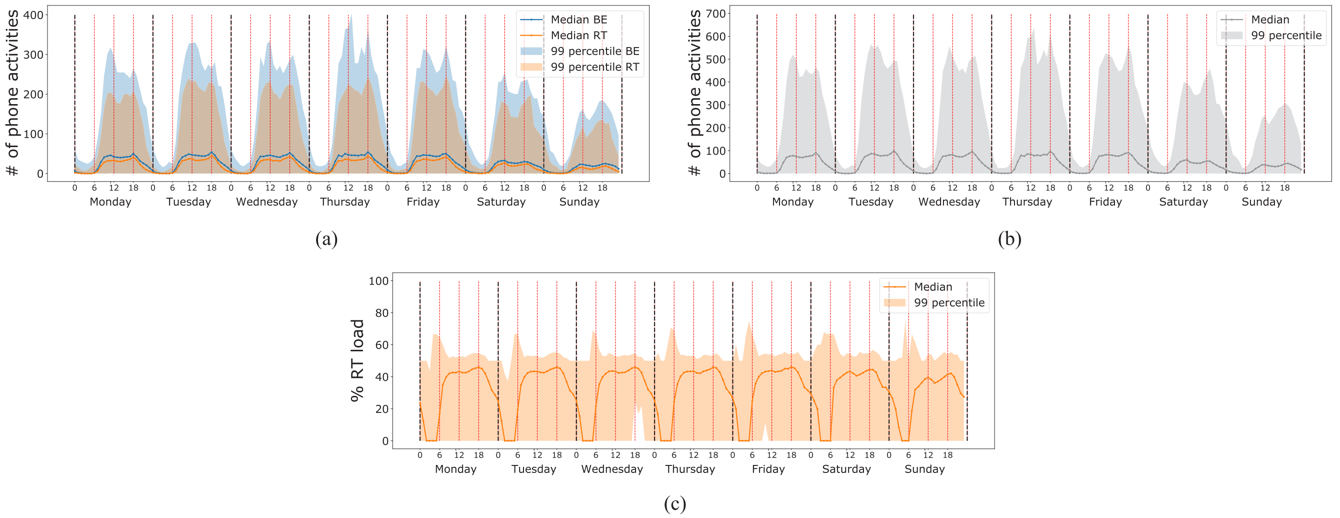
We consider three types of phone activities, namely, i.e., calls, missed calls and texts, in modeling the traffic of RT and BE slices. We regard missed calls[2] and text messages as part of the BE slice, whereas voice calls are responsible for RT slice traffic generation. According to [25], the delay budget for conversational voice traffic is 100 ms,[3] while it is up to 300 ms for TCP-based traffic.[4] To properly apportion the delay budgets of both the RT and BE slices, we observe that the figures that are provided by 3GPP for various types of traffic originate from considering the entire mobile operator network (RAN + backhaul + Core networks). In this study, we assume that the facilities are within the backhaul network, and we only aim at modeling the MEC part of the network slices; consequently, only a fraction of the delay budget that

[1] We use a radial basis function kernel with the target number of components set to 5.

[2] Missed calls rely on a call setup phase that does not require stringent delay constraints, as in the case of the session initiation protocol (SIP).

[3] The actual packet delays - especially for GBR traffic - should typically be lower than the delay that is specified for a QCI if the UE has sufficient radio channel quality [25].

[4] Conventionally, SMSs are delivered through an IP multimedia subsystem (IMS) that sets a budget delay of 100 ms for its operations. Nevertheless, in practice, the sensitivity of texts to the time delay is negligible; hence, we assign a delay budget of 300 ms to text messages, which is in accordance with TCP-based traffic.



(a)

(b)

(c)

FIGURE 6. Input instance: (a) the traffic profile for RT and BE slices, (b) the traffic profile for the single-slice case and (c) the pattern of the percentage of the RT slices traffic.

is reported by the standard should be considered. Therefore, after subtracting the RAN part (approximately 20 ms [25]), we obtain for the BE slice a delay budget that is approximately three/four times higher than that for the RT slice.

We exploit this result to generate an input instance of the optimization model that prioritizes the RT slice with respect to the BE slice and considers the network delay of the virtual links that are mapped onto the physical ones. Let us consider two virtual base stations, namely, i_{RT} and i_{BE}, that are deployed on physical base station i. We define their assignment costs to facility k, namely, $m_{i_{RT},k}$ and $m_{i_{BE},k}$, respectively, as follows: $m_{i_{BE},k} = p_{i,k}$ and $m_{i_{RT},k} = 4 \cdot p_{i,k}$.

3) MIGRATION COST

As discussed previously, changing the assignment of a vBS to a facility generates additional signaling traffic for reconfiguring the virtual link and for migrating the state-full MEC VNFs from the old to the new facility. In addition, the transfer of the VNF state from one facility to another increases the network delay. Accordingly, we approximate the migration cost between two facilities, namely, f and h, as the cost of the physical path that connects the two facilities: $l_{f,h} = g_{f,h}$.

In our experiments, the parameters α and β, which represent the relative importance of the assignment and migration costs, are fixed to 1.

D. REAL-TIME AND BEST-EFFORT TRAFFIC DEMAND

To create an input instance for the optimization model, we use a 1-hour time slot, which is a satisfactory compromise between a too-fine-grained sample (which could lead to many network reconfigurations) and a large time slot (which might lose the dynamics of the slice's traffic). For the traffic load of each slice, we compute the typical weekly median demand by aggregating the number of phone activities that are conducted in the same time slot of the week, in accordance with [36]. For each base station and for each slice, we compute the median load of the traffic that is generated in the same time slot by the phone activities that occurred in that base station. As a result, we obtain two time series for each base station, which represent the typical load patterns of the two slices across our dataset. In Fig. 6a, we report the load pattern of the typical weekly median demand; the lines represent the medians over all the base stations, whereas the areas represent the 99th percentile of the load distribution in each time slot. Both slices exhibit the typical aggregated pattern [37] with two peaks during the daytime, namely, at 12 p.m. and 6 p.m., and a limited amount of traffic during the night, with slight increases on Friday and Saturday nights. For completeness, in Fig. 6b, we present the load pattern of the single-slice case. Comparing the two patterns, they are similar in almost all time slots except for the two daily peaks; the highest peak always occurs at approximately 6 p.m. for the RT slice, while it occurs mainly at 12 p.m. for the other slice. Fig. 6c shows the percentage of the overall RT slice traffic with respect to the BE slice traffic. The line is the median of the percentage of RT traffic over time, while the area represents the 99th percentile of the distribution in each time slot. The peak of the median percentage of the RT slice is approximately 46% at 6 p.m. almost every day of the week. We use this information to set the percentage of the facility capacity that can be allocated to the traffic of the RT slice: $C_k^{RT} = 0.46 \cdot C_k$.

E. CAPACITY PARAMETER OF THE FACILITIES

The capacity of the facilities is a critical parameter of the optimization model. Its value may affect the capability of a facility to tolerate the traffic dynamics, which is henceforth referred to as the resilience, and the quality of the assignment plans. As described in Section III-B, the capacity of the facility is a constraint of the model; thus, the optimization algorithm tends toward using the entire available capacity of a facility if such a choice improves the value of the objective function. Inevitably, a slight increase in the traffic demand at a base station for which the associated facility exhausted its capacity will overload the facility and reduce the QoS. To overcome the problem, we consider two phases: In the first phase, namely, the *optimization phase*, we assess the level of QoS-net by using the traffic demand pattern that was computed in Section IV-D and adopting various capacity level settings C_k. Under these conditions, when a large capacity setting is utilized, more base stations are associated with their optimal facility. In contrast, a small capacity setting forces the optimization algorithm to provide less optimal but more robust assignments. In the second phase, namely, the *evaluation phase*, we assess the network resilience under practical traffic conditions for each slice. In this phase, the capacity values represent the levels that the operators wish to evaluate. They are assumed to be equal to or greater than the settings that were applied during the optimization phase. The gap between the two levels of capacity corresponds to the additional traffic that the facilities can tolerate without reducing the QoS-load. The choice of which capacity pair to adopt should be guided by three factors: (i) the suitable trade-off between the two components of QoS, (ii) the degree of resilience to traffic dynamics and (iii) the CAPEX/OPEX budget, which limits the maximum value of the capacity in the evaluation phase.

We compute the minimum required capacity for managing all the traffic in each time slot for the optimization instance; this capacity value represents the minimum capacity that renders the optimized solution feasible. In the following, we will denote it as *MIN*. This setting is not suitable for a practical deployment because it is too sensitive to traffic variations; however, it is a satisfactory indicator of the minimum required investment and of the basic performance to pursue. In addition, we set another capacity value, which is denoted as `MAX` and corresponds to the capacity that is required by each facility for dealing with the maximum traffic load the network can experience. This latter value leads to an over-provisioning of resources, but it will protect the mobile operator against service disruptions. In accordance with the data in our dataset, the `MAX` value is approximately 50% higher than the `MIN` value. The large distance between the two values is mainly due to the occurrence, within the time frame that is covered by our dataset, of an extraordinary event (an earthquake in the north of Italy) that caused an exceptional increase in the traffic demand. If excluding this event, the gap falls to approximately 12.5%. In our analysis, we identify five additional intermediate values of the additional capacity up to the maximum of 25%, namely, `MIN+5%`, `MIN+10%`, `MIN+15%`, `MIN+20%`, and `MIN+25%`, for tolerating an additional demand of up to twice that observed in the dataset.

The next sections will present a thorough analysis of a wide set of parameter combinations and will highlight the benefits and drawbacks of the choices.

V. ASSIGNMENT PLAN EVALUATION

In the following, the label *Baseline* denotes the model that is described in [15], which does not consider the slice organization, whereas the label *Multislice* denotes the new model.

A. TEMPORAL ANALYSIS

During the optimization phase, the network operator aims at realizing the optimal trade-off between two highly interwoven objectives, namely, the maximization of QoS – for the RT slice – and the maximization of the additional capacity that will be available during the evaluation phase. In accordance with the arguments in Section IV-E, an effective balance between the expected QoS and the marginal available capacity gap at the facilities should be pursued. For a specified capacity value, a basic indicator of the suitability of QoS-net is the number of vBSs that are assigned to the optimal facility in each time slot. In Fig. 7, we report the percentage of vBSs that are associated with the optimal facility (on the left side) and the differences between the two models (on the right) for the RT slice. The *Multislice* model outperforms the *Baseline* for various capacity values. This is more evident when small capacities are used, e.g., MIN and MIN +5%, while the difference is less pronounced in the case of large capacities. Furthermore, the gain is not uniform throughout the time frame of a week; four separate levels of performance are clearly identified according to the day of the week and the part of the day: (i) In the morning (6 - 10 a.m.) and at night (8 p.m. - 12 a.m.) on weekdays, the gain is large. In these time periods, the *Multislice* model assigns an additional 20% of the vBSs to their best facility; (ii) the gain is slightly reduced, but still remarkable, during the peak hours (excluding 6 p.m.) on weekdays; (iii) the two models perform almost equally during the weekend due to the reduced amount of traffic; and (iv) the gain is slightly negative at 6 p.m. on working days, when a peak in the RT traffic demand regularly occurs (see Fig. 6a), which is followed by a significant decrease in the traffic demand in the next time slot. This represents a challenging combination for the model because the search for the best trade-off between the assignment and migration costs is pushed to the limit, and this becomes especially critical for the *Multislice* model, which prioritizes the RT traffic. To explain this, in Fig. 8a, we report the number of planned migrations[5] for the RT slice between each pair of consecutive time slots for both models (we only report the case with `MIN+5%` as the optimization phase capacity). The number of planned migrations remains almost comparable throughout the time slots of all workdays, except for the time slot of 6 p.m., when the number of migrations for the *Multislice* model is almost twice the number that are planned by the other model. The trade-off produces a slight penalization in terms of association costs at 6 p.m. (see Fig. 8b); however, this leads to a global benefit in terms of the total number of migrations.

[5] In this analysis, we consider a case of vBS migration in which a vBS is associated with a disjoint set of facilities in two consecutive time slots.

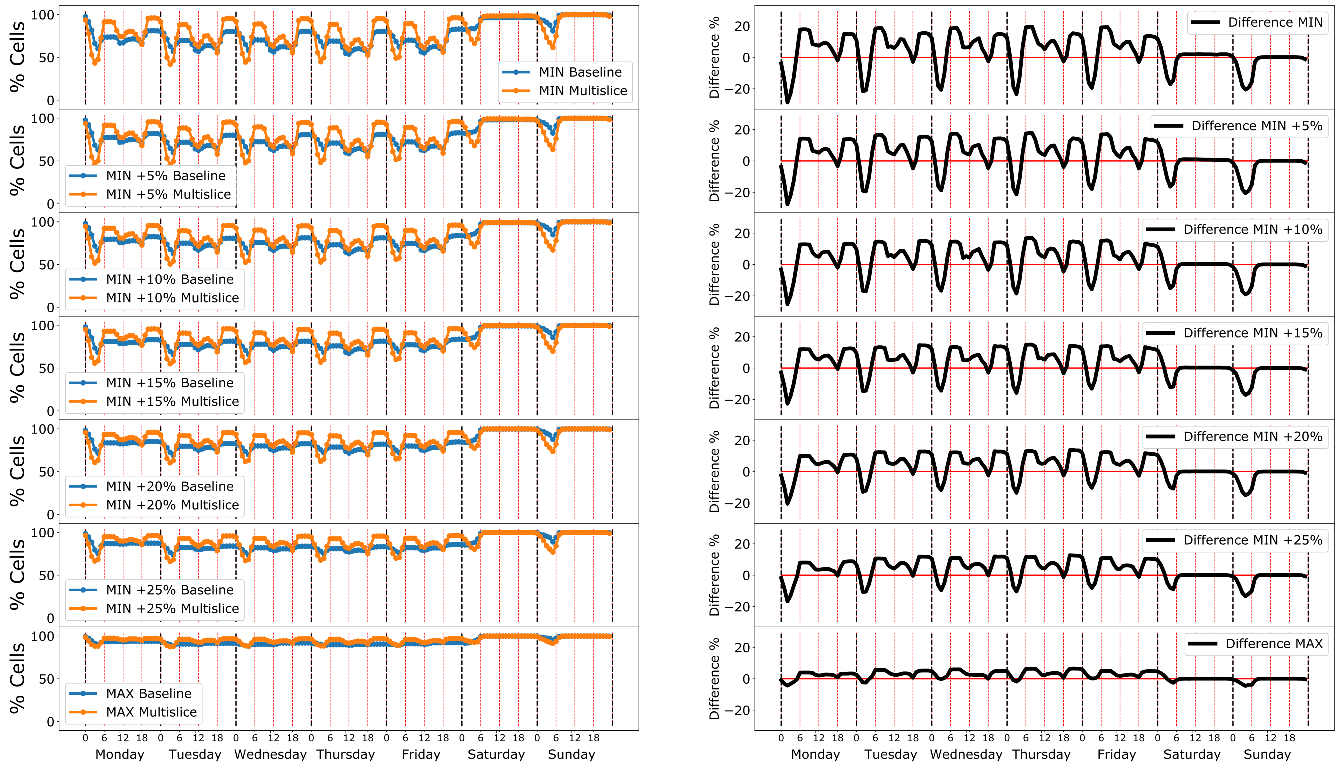
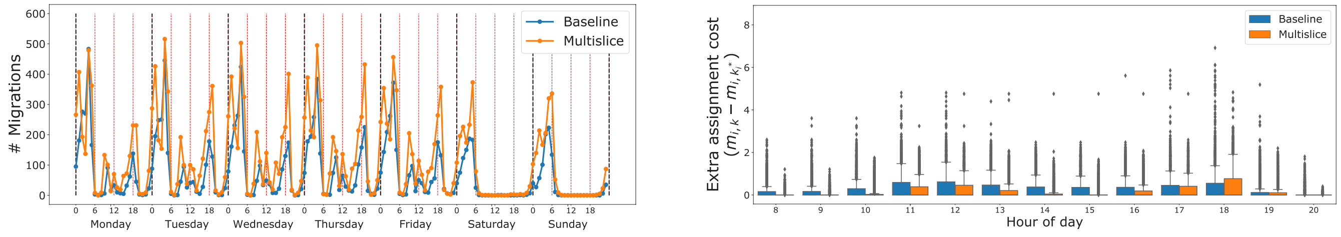


FIGURE 7. Percentage of vBSs that are associated with the optimal facility in each time slot for the RT slice (on the left) and the percentage difference between the two models (on the right).



(a)

(b)

FIGURE 8. (a) The number of planned migrations in each time slot for the RT slice. (b) A boxplot of the additional association cost ($m_{i,k} - m_{i,k_i^*}$) between 8 a.m. and 8 p.m. on Thursday using `MIN+5%` as the optimization phase capacity.

B. SPATIAL ANALYSIS

Additional information for a mobile operator is the identification of the base station and the time slots in which a poor QoS-net is provided. During the optimization phase, the analysis of the map of vBS-facility assignment provides insights regarding the occurrence of critical conditions in various areas of a city. We conduct this analysis for three time slots, namely, 8 a.m., 12 p.m., and 6 p.m., and we present the results for Thursday (the results for other weekdays are similar). These time slots are all significant from the traffic perspective. Fig. 9 shows maps of the traffic load for the RT slice in the three time slots, where the brighter the color, the higher the traffic load. Fig. 10 shows maps of the mean[6] difference between $m_{i,k}$ (the assignment cost between the vBS and the assigned facility) and $m_{i,k*}$ (the assignment cost between the vBS and the optimal facility) and enables the comparison of the two models under various capacity settings. The color of the area varies from a vivid green, when the vBS is optimally assigned to a facility, to a dark red, when the vBS is poorly assigned to a facility or a set of facilities.

At first glance, the *Multislice* model outperforms the *Baseline* model. This is especially evident in conditions of mid and low traffic, namely, at 8 a.m.. As shown in Fig. 10, the *Multislice* model can associate almost all vBSs to their best facility regardless of the optimization phase capacity level, whereas the *Baseline* model encounters minor difficulties when limited capacities are available (i.e., `MIN` and `MIN+5%`), for example, in sectors `C4` and `D4`. As the available capacity at the facilities increases, the gap between the two models decreases, but the *Multislice* model consistently

[6] Each difference is weighted by the percentage of traffic that is assigned to each vBS-facility pair by the fractional assignment.

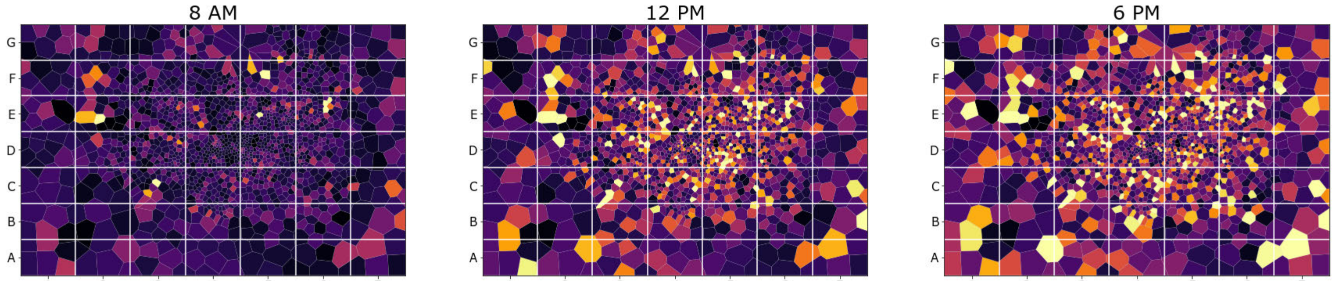


FIGURE 9. Map of the RT load from each vBS at 8 a.m., 12 p.m. and 6 p.m.; the brighter the color, the higher the traffic demand (the colors are not distinguishable in a grayscale print).

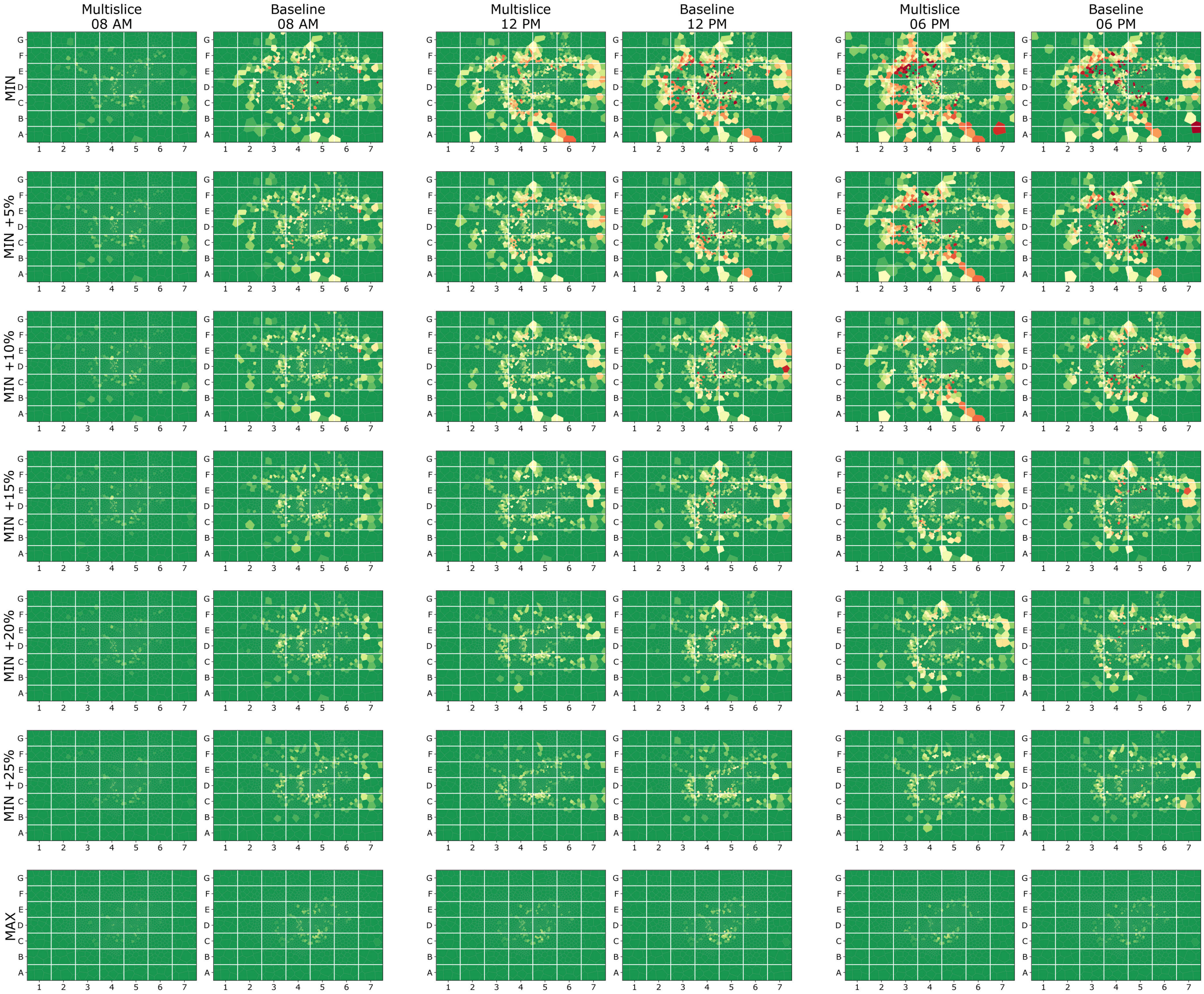


FIGURE 10. Comparison of the maps of the QoS-net for the RT slice at 8 a.m., 12 p.m. and 6 p.m. for each optimization phase capacity. Vivid green indicates that the vBS is optimally assigned, while full red indicates that the vBS is poorly assigned to a facility or a set of facilities (the colors are not distinguishable in a grayscale print).

provides the best assignments in the considered optimization phase scenarios.

As shown in Fig. 6a, the 12 p.m. and 6 p.m. time slots correspond to the daily peaks of RT traffic. The obtained assignments at 12 p.m. show that the *Multislice* model can avoid critical assignments, even when limited resources are available at the facilities (`MIN` and `MIN+5%`). Although the *Baseline* model can provide fair assignments for vBSs with the highest load of traffic (E2, D4 and D5 in Fig. 9), it fails to guarantee sufficient QoS-net to vBSs with mid/mid-high traffic load (the dark red areas in sectors C3, D3, E3 and D4 of Fig. 10). Starting from a MIN+10% capacity level, the *Baseline* model can provide a fair QoS-net level to the vBSs.

The resulting assignments at 6 p.m. show that both models encounter difficulties in ensuring fair vBS-facility association when the available capacity is highly limited, namely, MIN. When the smallest optimization phase capacity is used, the *Multislice* model is unable to offer a fair QoS to most vBSs in sectors E3, F3, D3, and C4 and to the large and heavily loaded cell in sector A7. In contrast, the *Baseline* model encounters more difficulties in sector C5, where a large set of vBSs are poorly assigned. However, even though the overall number of unfairly associated vBSs exceeds that for the *Multislice* model, they are scattered across the city. This enables the adoption of traditional load balancing algorithms to relocate users with poor QoS to their neighboring vBSs.

TABLE 3. Descriptive statistics of the percentage of the used capacity.

Model	Median	Mean	Std.
Baseline MIN +5%	95.87	80.84	26.08
Multislice MIN +5%	99.91	77.68	30.50
Baseline MIN +15%	88.72	73.82	29.65
Multislice MIN +15%	89.51	70.94	32.80

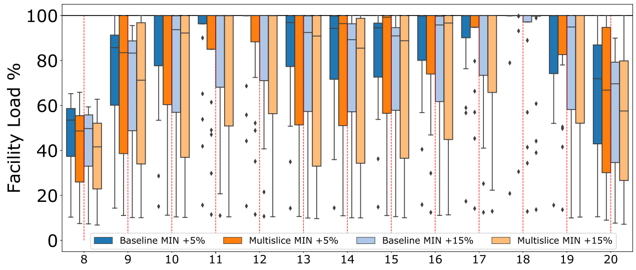


FIGURE 11. Facility load between 8 a.m. and 8 p.m. on Thursday of the optimization input week.

C. FACILITY LOAD

From a facility perspective, it is interesting to analyze the planned use of resources with the objective of providing the mobile operator with valuable temporal and spatial information regarding the potential criticality of the infrastructure. For the *Multislice* model, we compute the utilization of the RT capacity[7] of each facility, whereas for the *Baseline* model, we consider the entire facility capacity. In Fig. 11, we present a boxplot of the distribution of the used capacity between 8 a.m. and 8 p.m. on Thursday of the optimization input week (similar results are obtained on other weekdays) by choosing two capacity levels: MIN+5% and MIN+15%. The two models use the facility capacity differently. The *Baseline* model produces assignments that lead to a more homogeneous usage of all the facilities compared to the *Multislice* model. In contrast, the *Multislice* model, due to the higher assignment cost of the RT slice traffic, tends to change the balance of the facility utilization in favor of a better assignment. These observations are supported by the descriptive statistics that are reported in Table 3, according to which the mean of the percentage of the used capacities between 8 a.m. and 8 p.m. is lower for the *Multislice* model, while both the standard deviation and median are higher. These results demonstrate that the *Multislice* model provides a better assignment by exploiting all the available capacity of a small subset of facilities, while it substantially underutilizes the others.

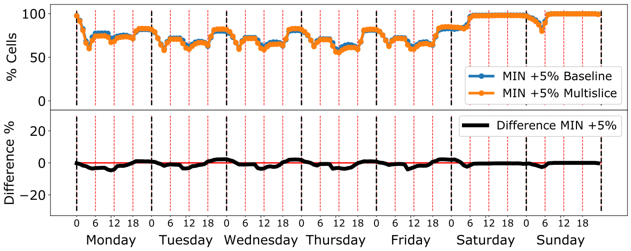


FIGURE 12. (Top) The percentage of vBSs that are associated with the best facility in each time slot for the BE slice in the MIN+5% scenario. (Bottom) The percentage difference between the two models.

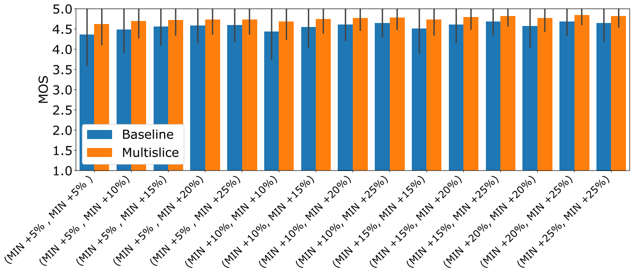


FIGURE 13. Mean and standard deviation of the MOS index of the QoS for all evaluation phase scenarios.

D. BE SLICE

The priority that is granted to the RT slice should not negatively affect the BE slice. In Fig. 12, for the BE slice with capacity level MIN+5%, we report the percentage of vBSs that are assigned to their optimal facility (top curve) and the difference between the two models (bottom). The figure clearly shows similar results for both models.

VI. QOS EVALUATION

Mobile operators can use the evaluation phase to obtain precise insights regarding the realizable level of QoS for specified infrastructure expenditures (i.e., CAPEX/OPEX). In our analysis, the CAPEX/OPEX is assumed to be a function of the facilities' capacity, namely, the larger the capacity, the larger the investments.

In the following, we consider various combinations of evaluation/optimization phase capacities. We consider only

[7] Recall that 46% of the capacity of each facility is preemptively allocated to the RT slice traffic.

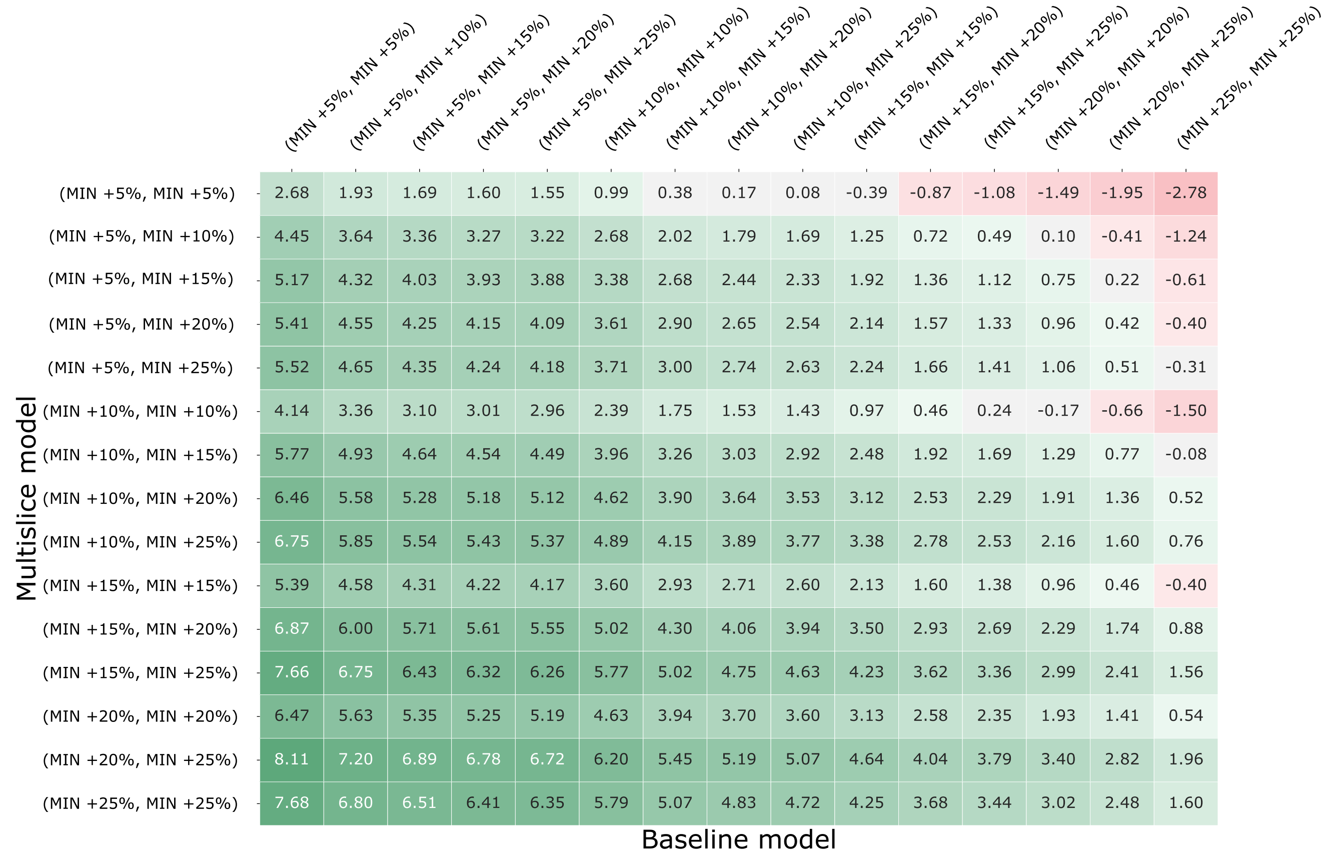


FIGURE 14. Comparison between the two models in various scenarios. The rows correspond to the *Multislice* model scenarios; the columns correspond to the *Baseline* model scenarios. Each cell of the matrix contains the average QoS percentage gain of the *Multislice* scenario with respect to the *Baseline* model.

the pairs in which the evaluation phase capacity is larger than or equal to the optimization phase capacity. In addition, we consider only the capacity levels between `MIN 5%` and `MIN 25%`, and we exclude the `MIN` and `MAX` levels because the former is too restrictive and unfeasible in a practical deployment and the latter is not interesting from the optimization perspective. In the following, we specify the pairs of capacities using the ordered pair notation (C_{opt}, C_{ev}), where C_{opt} is the level of the capacity that is used in the optimization phase, whereas C_{ev} is adopted in the evaluation phase, e.g., `(MIN +5%, MIN +20%)`. Finally, in this section, we focus only on the results regarding the RT slice, as both models perform similarly on the BE slice.

A. GENERAL QOS COMPARISON

The evaluation of QoS is conducted throughout the nine weeks that are covered by our dataset. For each pair of optimization and evaluation phase capacities and for each time slot, we compute the QoS. In Fig. 13, we report the mean and the standard deviation of the QoSs (by using the MOS index) that are realized by the two models in all the considered scenarios. The *Multislice* model can always ensure a higher and more stable QoS. The mean QoS value ranges between 4.62 and 4.84 in the *Multislice* model, while it is between 4.36 and 4.68 in the *Baseline* model. Moreover, the values of the standard deviation are lower for the *Multislice* model (between 0.22 and 0.50) than for the *Baseline* model (between 0.31 and 0.75).

In Fig. 14, we report the percentage difference of the QoSs that are realized by the two models in various capacity scenarios. The rows correspond to the optimization/evaluation phase capacity pairs for the *Multislice* model, while the columns correspond to those for the *Baseline* model. Each cell contains the average QoS percentage gain that is realized in the *Multislice* scenario with respect to the *Baseline* one. The diagonal provides a comparison of the two models in the same scenario. By reading the rows of the matrix, we assess the gain that the *Multislice* model can realize in a scenario with respect to various scenarios of the *Baseline* approach. The upper-triangular part of the matrix provides insights regarding the capacity settings in which the *Multislice* model has fewer resources available than the *Baseline* model. When evaluating the two models under the same evaluation scenario (the diagonal), the larger the gap between the optimization and evaluation phase capacities, the larger the gain. The scenarios with the largest gain,

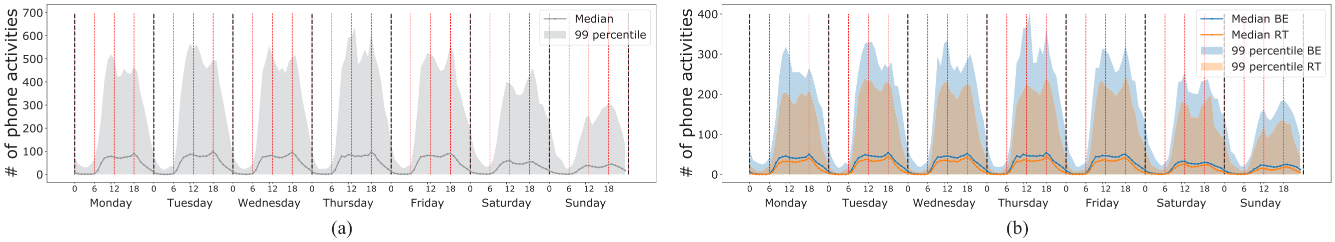


FIGURE 15. Evaluation phase week: (a) the traffic profile for the single-slice case and (b) the traffic profile for the RT and BE slices.

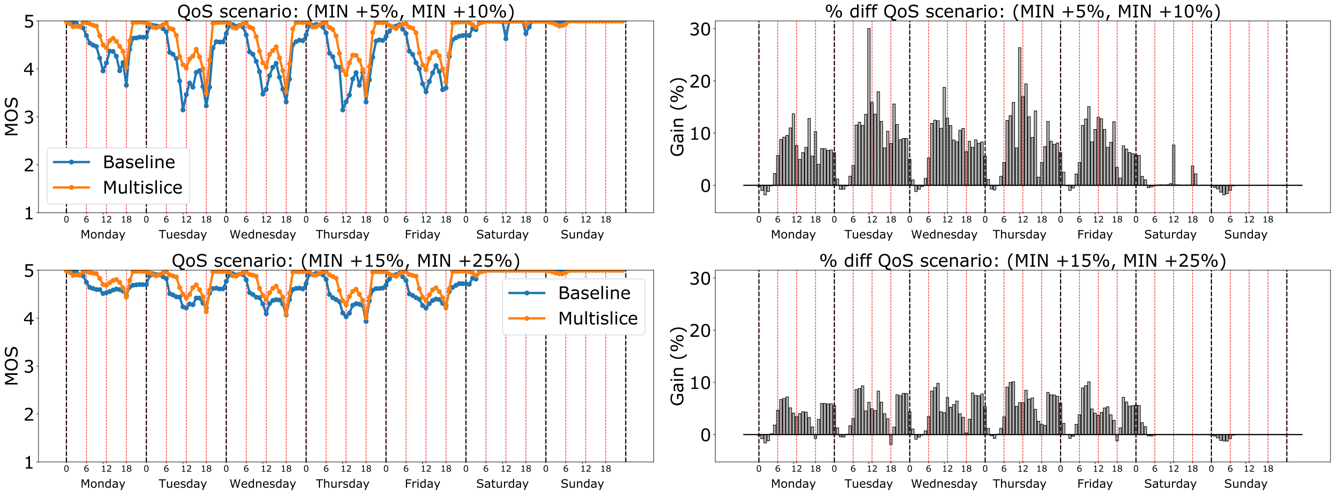


FIGURE 16. Weekly mean level of QoS (on the left) and the percentage of the gain between the two models (on the right).

e.g., (MIN +5%, MIN +20%) and (MIN +5%, MIN +25%), represent the conditions in which the operator has been conservative (in terms of capacity) during the optimization phase and liberal in the deployment. In such a condition, the *Multislice* model benefits from the prioritization of the RT slice, which provides higher QoS-net without degradation of QoS-load.

Fig. 14 clearly shows that the *Multislice* model ensures higher QoS even when the *Baseline* model has more resources. For example, the *Multislice* model in scenario (MIN +5%, MIN +5%) performs as well as the *Baseline* model in scenario (MIN +10%, MIN +25%). When the optimization and evaluation phase capacities are equal, the gain of the *Multislice* model decreases as the facilities are overloaded, thereby causing a degradation of the QoS-load.

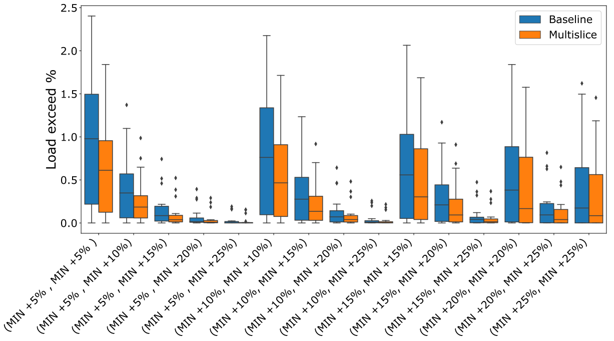


FIGURE 17. Boxplot of the distributions of the capacity violation index for both models in every evaluation phase scenario.

B. WEEKLY QOS PATTERN

The previous section provides an overview of the performances of the two models throughout the time period that is covered by the dataset. Here, we focus on the QoS that can be realized at each time slot. The analysis is conducted by considering a single week and two scenarios: (MIN +5%, MIN +10%) and (MIN +15%, MIN +25%). The former combines a conservative choice during the optimization phase and a limited amount of resources during the deployment, while the latter corresponds to a less cautious optimization phase and high resource availability during the evaluation phase.

To conduct the analysis, we select the traffic load of the last week in the dataset, as its temporal pattern is highly similar to those of the other evaluation phase weeks. In Fig. 15a, we report the median and the 99th percentile of the base station's traffic demand in the case of a single slice, while in Fig. 15b, we report the traffic demand for the scenario with two slices. The single-slice traffic pattern shows a peak of

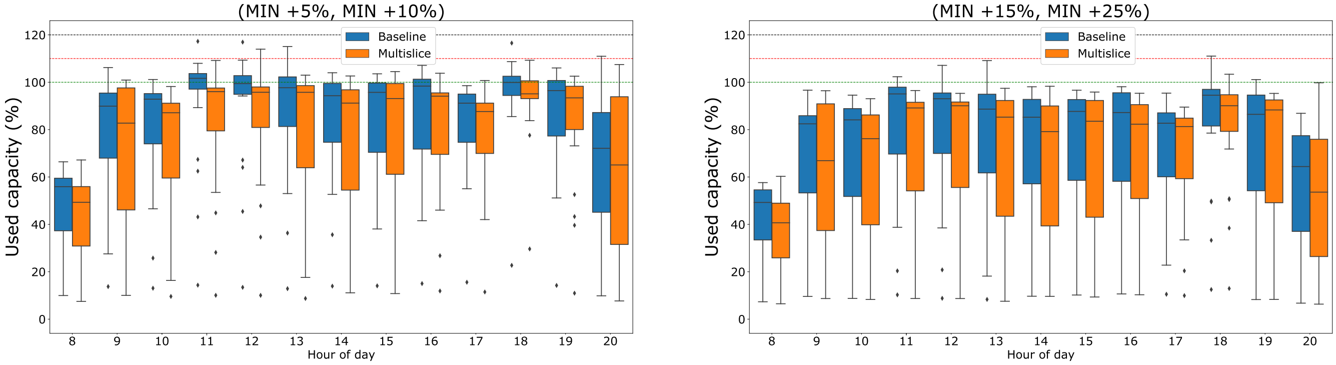


FIGURE 18. **Facility load between 8 a.m. and 8 p.m. on Thursday of the evaluation phase week for scenarios** (MIN +5%, MIN +10%) **(on the right) and** (MIN +15%, MIN +25%) **(on the left).**

traffic on Thursday at 1 p.m.; however, if we consider the two slices separately, we observe that this is the peak of the BE slice, while the peak of the RT slice remains at 6 p.m., as in the optimization input week.

In Fig. 16, we report the weekly mean level of QoS (on the left), together with the percentage of the gain between the two models (on the right). The results demonstrate that the *Multislice* model provides a higher QoS level in almost every time slot and under the highest traffic conditions (the *Baseline* model performs slightly better only late at night). In scenario (MIN +5%, MIN +10%), the gain increases up to 25-30% around the peak traffic demand in the single-slice setting (see Fig. 15a). Due to the compromise between the migration and the assignment costs (see Section V-A), every day at 6 p.m., we observe a slight decrease in the gain in scenario (MIN +15%, MIN +25%); however, this is a negligible issue because the loss is highly marginal and it is highly predictable even during the optimization phase.

C. FACILITY LOAD

The capacity at the facilities is never exceeded during the optimization phase due to the model constraint. As shown in Section V-C, both models exploit the entire available capacity to improve the QoS-net. This aspect is emphasized in the case of the *Multislice* model, which tends to use all the available capacity in a small subset of facilities. Consequently, during the evaluation phase, a change in the traffic demand with respect to the optimization input week could lead to a degradation of the QoS-load due to capacity violation. We analyze the condition of facility overload by measuring the mean additional capacity for each facility that is used throughout the nine weeks. Formally, we measure the capacity violation using the following index:

$$\frac{\sum_{t\in T}\max\{\sum_{i\in A} w_i^t x_{i,k}^t - C_k, 0\}}{C_k \cdot |T|} \tag{17}$$

where T is the set of the time slots, C_k is the capacity of facility k, w_i^t is the traffic demand of vBS i at time t, and $x_{i,k}^t$ is the percentage of the traffic demand of vBS i that is assigned to facility k. We use the max operator to ensure that the numerator always non-negative since we are interested in capacity violation only.

In Fig. 17, we present a boxplot of the distributions of the capacity violation index for both models in every evaluation phase scenario. The results demonstrate that the *Multislice* model is more resilient than the *Baseline* approach in all scenarios, especially in those with no or limited capacity gap between the optimization and evaluation phase capacity levels. The larger the gap, the smaller the percentage by which the capacity is exceeded, e.g., (MIN +5%, MIN +25%) and (MIN +10%, MIN +25%), because the available capacity is sufficiently large for handling the increase in the traffic demand throughout the evaluation phase weeks.

The above results provide a broad outline of the facility overload. Now, we deepen the analysis by evaluating the capacity use on an hourly basis with the objective of identifying the most critical time slots. In Fig. 18, we report the percentages of the used capacity between 8 a.m. and 8 p.m. on Thursday for two evaluation phase scenarios: (MIN +5%, MIN +10%) and (MIN +15%, MIN +25%). In both scenarios, the two models exhibit the same patterns (see Fig. 11). The *Baseline* model distributes the load more fairly across the facilities, while the *Multislice* model assigns most of the traffic to a small set of facilities. Nevertheless, due to the differences in the temporal dynamics of traffic demand between the two slices (see Fig. 15a and 15b), the *Baseline* model must deal with the peak of the single-slice traffic demand approximately 12 p.m., which causes the overload of many facilities (almost 50% in the (MIN +5%, MIN +10%) scenario). In contrast, the *Multislice* model can mitigate the overload conditions, e.g., only 25% of the facilities are overloaded at 6 p.m. in the (MIN +5%, MIN +10%) scenario. The differences in the results are also due to the differences in the traffic dynamics of the two network slices during the evaluation phase week. We observed that the BE slice traffic demand increases by up to 17-18% with respect to the optimization input week, while the traffic demand increases by up to 12-14% in the case of the RT slice.

The *Baseline* model is unable to distinguish between these two flow characteristics and suffers from poor QoS-load, especially with a limited amount of resources.

VII. CONCLUSION

In this paper, we address the problem of proactively planning the BS – MEC facility associations in multi-slice scenarios. We consider two slices, namely, RT and BE, which are modeled by using an anonymized mobile phone dataset. The results demonstrate that by decoupling the RT and BE traffic demands, the mobile operator can improve the base station-facility assignments and ensure superior quality of service provisioning, even when limited resources are available. The results demonstrate that there is only one critical condition, namely, at 6 p.m. on weekdays and with a highly limited capacity level, in which the *Multislice* model performs slightly worse than the *Baseline* model. Nevertheless, this condition does not represent a critical issue for operators since (i) the bottom line capacity is unlikely to be used in practical deployments, (ii) the event is highly predictable, thereby enabling the adoption of a tailored network configuration in those time slots, and (iii) the performance substantially improves when a slightly larger capacity is adopted. Interestingly, the *Multislice* model realizes higher QoS even when it operates with fewer resources than those assigned to the *Baseline* model.

ACKNOWLEDGMENT

This work was funded by the Italian Ministry for Instruction, University and Research under the PON VASARI (ARS01_00456) Project.

REFERENCES

[1] M. Condoluci and T. Mahmoodi, "Softwarization and virtualization in 5G mobile networks: Benefits, trends and challenges," *Comput. Netw.*, vol. 146, pp. 65–84, Dec. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128618302500

[2] *Multi-Access Edge Computing (MEC); Framework and Reference Architecture*, Standard ETSI GS MEC 003, Jan. 2019.

[3] *Network Functions Virtualisation (NFV); Infrastructure Overview*, Standard ETSI GS NFV-INF 001, 2013.

[4] *OSM Release FIVE Technical Overview*, ETSI, Sophia Antipolis, France, 2019.

[5] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128619304773

[6] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 3, pp. 462–476, Sep. 2016.

[7] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 124–161, 1st Quart., 2016.

[8] R. A. Addad, M. Bagaa, T. Taleb, D. L. C. Dutra, and H. Flinck, "Optimization model for cross-domain network slices in 5G networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 5, pp. 1156–1169, May 2020.

[9] P. T. A. Quang, A. Bradai, K. D. Singh, G. Picard, and R. Riggio, "Single and multi-domain adaptive allocation algorithms for VNF forwarding graph embedding," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 1, pp. 98–112, Mar. 2019.

[10] Q. Zhang, X. Wang, I. Kim, P. Palacharla, and T. Ikeuchi, "Vertex-centric computation of service function chains in multi-domain networks," in *Proc. IEEE NetSoft Conf. Workshops (NetSoft)*, Jun. 2016, pp. 211–218.

[11] R. Mijumbi, J. Serrat, J.-L. Gorricho, J. Rubio-Loyola, and S. Davy, "Server placement and assignment in virtualized radio access networks," in *Proc. 11th Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2015, pp. 398–401.

[12] 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.

[13] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant mobile networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, Apr. 2019.

[14] R. Li, Z. Zhao, Q. Sun, C.-L. I, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.

[15] A. Ceselli, M. Fiore, M. Premoli, and S. Secci, "Optimized assignment patterns in mobile edge cloud networks," *Comput. Oper. Res.*, vol. 106, pp. 246–259, Jun. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305054818300595

[16] *Telecommunication Management; Study on Management and Orchestration of Network Slicing for Next Generation Network*, document TS 28.801, Release 15, 3GPP, Jan. 2018.

[17] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804516301989

[18] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should I slice my network?: A multi-service empirical evaluation of resource sharing efficiency," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2018, pp. 191–206. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241567

[19] J. Gil Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 3, pp. 518–532, Sep. 2016.

[20] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[21] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-Art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 416–464, 1st Quart., 2018.

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

[23] M. F. Bari, R. Boutaba, R. Esteves, L. Z. Granville, M. Podlesny, M. G. Rabbani, Q. Zhang, and M. F. Zhani, "Data center network virtualization: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 909–928, 2nd Quart., 2013.

[24] S. Wang, J. Xu, N. Zhang, and Y. Liu, "A survey on service migration in mobile edge computing," *IEEE Access*, vol. 6, pp. 23511–23528, 2018.

[25] *Policy and Charging Control Architecture*, document TS 23.203, Release 15, 3GPP, Sep. 2018.

[26] K. Han, S. Li, S. Tang, H. Huang, S. Zhao, G. Fu, and Z. Zhu, "Application-driven end-to-end slicing: When wireless network virtualization orchestrates with NFV-based mobile edge computing," *IEEE Access*, vol. 6, pp. 26567–26577, 2018.

[27] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[28] L. Tian, Y. Zhou, Y. Wang, J. Yang, Q. Sun, J. Yuan, and B. Yang, "Evaluation methodology for virtual base station platforms in radio access networks," *IEEE Access*, vol. 6, pp. 49366–49374, 2018.

[29] M. Conforti, G. Cornuéjols, and G. Zambelli, Eds., *Integer Programming*. Cham, Switzerland: Springer, 2014.

[30] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[31] *IBM Optimizer Webpage*. Accessed: Apr. 2019. [Online]. Available: https://www.ibm.com/analytics/cplex-optimizer

[32] G. Desaulniers, J. Desrosiers, and M. M. Solomon, Eds., *Column Generation*. New York, NY, USA: Springer, 2005.

[33] G. L. Nemhauser, "Column generation for linear and integer programming," *Optim. Stories*, vol. 20, p. 64, Aug. 2012.

[34] J. Martin-Perez, L. Cominardi, C. J. Bernardos, A. de la Oliva, and A. Azcorra, "Modeling mobile edge computing deployments for low latency multimedia services," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 464–474, Jun. 2019.

[35] C. Quadri, M. Zignani, L. Capra, S. Gaito, and G. P. Rossi, "Multidimensional human dynamics in mobile phone communications," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e103183, doi: 10.1371/journal.pone.0103183.

[36] A. Ceselli, M. Fiore, A. Furno, M. Premoli, S. Secci, and R. Stanica, "Prescriptive analytics for MEC orchestration," in *Proc. IFIP Netw. Conf. (IFIP Netw.) Workshops*, Zürich, Switzerland, May 2018, pp. 1–9. [Online]. Available: https://hal.sorbonne-universite.fr/hal-01740816

[37] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 1147–1161, Apr. 2017.

CHRISTIAN QUADRI (Member, IEEE) received the Ph.D. degree in computer science from the University of Milan, in 2015. His current research interests focus on 5G communication networks and MEC/FOG-computing and span the following aspects of anticipatory networking, (i) extraction of spatiotemporal behaviors of mobile users via the analysis of mobile network datasets, (ii) mining and prediction of network/service resource utilization, and (iii) optimal placement of network/service resources in mobile operator networks (core and edge). His previous research focused on the design and evaluation of routing protocols for opportunistic networks and the analysis of communication patterns and social interactions of mobile users. He is a currently Postdoctoral Researcher with the Computer Science Department, University of Milan.

MARCO PREMOLI is currently a Postdoctoral Researcher with the Department of Computer Science, University of Milan. His current research interests are optimization algorithms and decision support systems in the fields of smart manufacturing and telecommunication networks.

ALBERTO CESELLI is currently an Associate Professor in computer science with the Department of Computer Science, University of Milan. His research interests include prescriptive data analytics, mathematical programming, computational integer programming, and design and experimental analysis of algorithms for combinatorial optimization problems.

SABRINA GAITO (Member, IEEE) received the degree in physics, the master's degree in material science, and the Ph.D. degree in applied mathematics from the University of Milan, Italy, in 1996, 1998, and 2002, respectively. She is currently an Associate Professor with the University of Milan, where she also teaches social media mining and computer networks. Her research activity takes place within both network science, with a focus on the application of complex network theory to social networking, human mobility and behaving, and network technology, with a focus on ad hoc networks and mobile applications.

GIAN PAOLO ROSSI (Member, IEEE) spent two Postdoctoral years at the Joint Research Center, EC, Ispra, where he has participated in the development and the deployment of the first European packet switched network (EIN). In 1980, he joined the University of Milano, as an Assistant Professor. From 2000 to 2006, he has chaired the Faculty Track in computer science. He has coauthored almost 100 scientific articles and coordinated/participated in several national and European research projects. His main research focus is on computer networks, including architecture and protocol design, performance evaluation and measurement, modeling, and analysis. His current research interests include social networks and network science.