Annals of Telecommunications
https://doi.org/10.1007/s12243-019-00722-w

Algorithms for the design of 5G networks with VNF-based Reusable Functional Blocks

Luca Chiaraviglio[1,2] · Fabio D'Andreagiovanni[3,4] · Simone Rossetti[1,2] · Giulio Sidoretti[1,2] · Nicola Blefari-Melazzi[1,2] · Stefano Salsano[1,2] · Carla-Fabiana Chiasserini[5,6] · Francesco Malandrino[6]

Received: 29 June 2018 / Accepted: 26 June 2019

Abstract

We face the problem of designing a 5G network composed of Virtual Network Function (VNF)-based entities, called Reusable Functional Blocks (RFBs). RFBs provide a high level of flexibility and scalability, which are recognized as core functions for the deployment of the forthcoming 5G technology. Moreover, the RFBs can be run on different HardWare (HW) and SoftWare (SW) execution environments located in 5G nodes, in line with the current trend of network softwarization. After overviewing the considered RFB-based 5G network architecture, we formulate the problem of minimizing the total costs of a 5G network composed of RFBs and physical 5G nodes. Since the presented problem is NP-Hard, we derive two algorithms, called SFDA and 5G-PCDA, to tackle it. We then consider a set of scenarios located in the city of San Francisco, where the positions of the users and the set of candidate sites to host 5G nodes have been derived from the WeFi app. Our results clearly show the trade-offs that emerge between (i) the total costs incurred by the installation of the 5G equipment, (ii) the percentage of users that are served, and (iii) the minimum downlink traffic provided to the users.

Keywords 5G networks · 5G Design · CAPEX reduction · 5G performance evaluation · Network softwarization

1 Introduction

According to the 5G Public Private Partnership (PPP), the forthcoming 5G technology is going to be a platform able to trigger new business models [1], involving the entry into the market of verticals, such as industries, manufacturing, and entertainment. In this scenario, the 5G network will be able to provide, among the other features, an extremely high bandwidth to users, with the deployment of the e-MBB (enhanced Mobile BroadBand) use case [1]. To achieve this goal, the network will extensively exploit the cloud concept, coupled with the need of slicing the physical resources into virtualized ones.

In this scenario, the softwarization paradigm is emerging as a promising candidate to realize future networks [2]. According to this trend, both the networking and computing functions are virtualized, and are thus decoupled from the underlying HW. More in detail, 5G will intensively exploit the deployment of virtual functions to realize both the core and the mobile network [3]. Thanks to the possibility of running virtual functions on shared HW, it becomes possible to deploy a flexible and scalable mobile network [4], able to guarantee extreme performance to users while reducing both the design and the maintenance costs. In this scenario, the Superfluidity (SF) project, funded by the European Commission through the Horizon 2020 Call, aims at providing superfluidity in the Internet, by instantiating services on-the-fly, run them at different network levels (i.e., core, aggregation, edge) and move them transparently

✉ Luca Chiaraviglio
luca.chiaraviglio@uniroma2.it

Fabio D'Andreagiovanni
d.andreagiovanni@hds.utc.fr

Carla-Fabiana Chiasserini
carla.chiasserini@polito.it

Francesco Malandrino
francesco.malandrinoi@eiit.cnr.it

[1] Consorzio Nazionale Interunivesitario per le Telecomunicazioni, Rome, Italy

[2] EE Department, University of Rome Tor Vergata, Rome, Italy

[3] National Center for Scientific Research (CNRS), Paris, France

[4] CNRS, Heudiasyc UMR 7253, Sorbonne Universités, Université de Technologie de Compiègne, CS 60319, 60203 Compiègne, France

[5] DET Department, Polytechnic Univerisity of Turin, Torino, Italy

[6] IEIIT National Research Center (CNR), Rome, Italy

Published online: 09 July 2019

to different 5G nodes. The core of the project is the definition of a cloud-based 5G converged solution, in which softwarized Virtual Network Function (VNF)–based components, called Reusable Functional Blocks (RFBs), are deployed [5]. More in detail, the RFBs implement all the required functionalities in the network, ranging from low-level ones (such as the Remote Radio Head–RRH) to high level tasks, thus matching the required level of flexibility and scalability of future 5G networks.

In this context, several questions are arising, such as: Is it possible to derive a model to minimize the installation costs of an RFB-based 5G architecture, while still guaranteeing the 5G service to users? How to design a set of smart algorithms to solve the considered problem? How to derive meaningful scenarios to test the proposed solutions? The answer to these questions is the goal of this paper. More in detail, our original contributions can be summarized as follows:

- we optimally formulate the problem of minimizing the installation costs of an RFB-based 5G network composed of different types of RFBs. Our formulation is able to produce as output the set of installed 5G nodes, the RFBs running on them, and the assignment of users to the RRH RFBs;
- we provide two efficient heuristics, called SuperFluid Design Algorithm (SFDA) and 5G Performance Clustered Design Algorithm (5G-PCDA), to solve the problem. While SFDA is tailored to the reduction of the installation costs, 5G-PCDA tends to efficiently maximize the number of served users;
- we consider a set of scenarios based on realistic measurements derived from the WeFi app [6];
- we run SFDA and 5G-PCDA on the considered scenarios, and we deeply analyze the trade-offs that emerge.

To the best of our knowledge, none of the previous works has conducted a similar analysis. The closest paper to our work is [7], in which the authors have targeted the efficient management of the RFBs in a 5G network, with the goal of maximizing the traffic per user or the number of used nodes. However, the work in [7] is tailored to the management phase, i.e., the design of the network is not considered at all, and in particular the costs that are incurred by the network owner from the installation of 5G nodes and RFBs are neglected. Moreover, in [7], the authors do not ensure a minimum traffic to users. Hence, a user may receive a very low amount of downlink traffic. To overcome these issues, in this work, we explicitly tackle the design phase of the network, in order to decide where to install the 5G nodes and where to place the RFBs. Moreover, we impose that users request a given amount of traffic, which has to be satisfied by the 5G network. As a result, the problem faced in this work is complementary to [7]. In particular, the elements installed during the design phase, which are selected by this work, can be used as input for the management one.

Actually, this work is an extended version of [8], where we preliminary investigate the design problem in an RFB-based 5G network. Differently from [8], in this work, we go four steps further by:

- showing that the considered problem is NP-Hard, and therefore very challenging to be solved apart from simple cases;
- designing the brand-new 5G-PCDA heuristic, which is able to efficiently solve the problem even for large instances;
- considering a new set of scenarios derived from realistic measurements from the city of San Francisco;
- running both SFDA and 5G-PCDA on the new scenarios, and deeply analyzing the trade-offs that emerge.

Our results clearly show that the costs for designing the RFB-based 5G network can be taken into account, while guaranteeing an adequate QoS perceived by users. Even though the results presented in this paper are promising, we point out that this work is a first step towards a more comprehensive approach, in which finer RFBs (smaller than the ones considered in this work) are used. In addition, another interesting research activity will be to take into account the users mobility, as well as considering the uncertainty of the users traffic. We leave the evaluation of these aspects as future work.

The rest of the paper is organized as follows. Section 2 overviews the related works. The RFB-based 5G architecture is described in Section 3. The optimal formulation is detailed in Section 4. Section 5 includes the description of the SFDA and 5G-PCDA algorithms. Section 6 details the scenarios and the parameter settings. The performance of the algorithms is evaluated in Section 7. Finally, conclusions are drawn in Section 8.

2 Related work

We briefly review the literature related to this work. More in depth, the basic concepts concerning the decomposition of the 5G services into a set of Virtual Network Functions (VNFs) are discussed in [3]. In addition, in [9], the author focuses on the concept of network function decomposition in conjunction with its relation to network slicing. Both [3] and [9] discuss the architectural aspects of the decomposition but do not provide an allocation model.

Several works have considered the problem of optimal placement of VNFs. In [10], the authors consider as VNF the Serving Gateway (SGW) and PDN Gateway (PGW) functions of the mobile core network. The proposed VNF placement model aims at minimizing the transport network load overhead against several parameters such as data-plane delay, number of potential datacenters and SDN control overhead. In [11], the considered VNFs are firewalls, load balancers, and VPN nodes. An Integer Linear Programming (ILP) model is proposed for the VNF placement and chaining problem. The set of PoPs on which it is possible to place the VNFs is given. In order to cope with large infrastructures, a heuristic procedure is proposed for efficiently guiding the ILP solver towards feasible, near-optimal solutions. In [12], the authors focus on a single centralized data center infrastructure and consider as a cost the utilization of the data center infrastructure. Two heuristic strategies for initial VNF deployment are compared. Finally, the authors of [13] study the influence of NFV on CAPital EXpenditure (CAPEX) of cloud-based networks and compare it with traditional implementation without NFV in different scenarios. However, no general optimization models are provided. In addition, none of these papers targets the radio access part of the network, which is instead taken into account by our work.

Focusing instead on the functionalities provided by the network, in [14], the authors propose a cloud-based wireless network architecture, which is composed of a mobile cloud, a Cloud Radio Access Network (CRAN), a mobile network, and a data center. In addition, in [15], the authors details a holistic architecture where Network Function Virtualization (NFV), Software Defined Radio (SDR), and Software Defined Networking (SDN) are exploited for the deployment of 4G/5G networks. Moreover, the challenges and the requirement for the adoption of dense 5G deployments and centralized processing are discussed in [16], highlighting the important role of cloud technologies and flexible functionality assignment. Although these works are prominent, they are mainly tailored to an architectural level, without considering the modeling of the problem or the definition of algorithms.

3 RFB-based 5G architecture

We report here a brief overview of the RFB-based 5G architecture, which is detailed in [5]. More in depth, the main building blocks of the architecture are represented by the Reusable Functional Blocks (RFBs), which are SoftWare (SW) functions realizing specific tasks. The RFBs are executed on the HardWare (HW) installed on the 5G nodes. One of the main advantages of such solution is the fact that the RFBs can be allocated and deallocated on the 5G nodes, in order, e.g., to satisfy the traffic spikes from users and/or to take into account the user mobility. In general, the RFB is a generalization of the Virtual Network Function (VNF) entity [17], which is able to run on different HW and SW execution environments. Eventually, the RFBs can be also decomposed in other RFBs, thus realizing less complex and/or recursive functions. We leave this last aspect as future work, while here we mainly focus on the design of a 5G architecture composed of standard RFBs.

Focusing on the tasks realized on the RFBs, we consider the following ones: (i) Remote Radio Head (RRH) RFB, (ii) Base Band Unit (BBU) RFB, and (iii) Mobile Edge Computing (MEC) RFB. More in detail, the RRH RFB is in charge of providing the physical signal to the users, by exploiting the Multi User Multiple Input Multiple Output (MU-MIMO) technology [18, 19]. On the other hand, the base band signal is managed by the BBU RFB, which acts as a middle layer between the physical level and the upper ones. Eventually, the computing functionalities, which, e.g., include the provisioning of a High-Definition (HD) video service to users, are realized by the MEC RFB. From a logical point of view, the RFBs are organized in chains. In this work, we consider a logical chain in which each RRH RFB is connected to a BBU RFB, which is in turn linked to a MEC RFB.

The RFBs are then run on the HW provided by the 5G nodes. More in detail, each 5G node is able to host the RRH RFB and the low-level functions of the BBU RFB on a Dedicated HardWare (DHW), while the high level functions of the BBU RFB and the MEC RFB are run on the Commodity HardWare (CHW). The RRH RFB is then connected to a set of physical antennas, which cover an area including the users. Figure 1 reports a scheme of

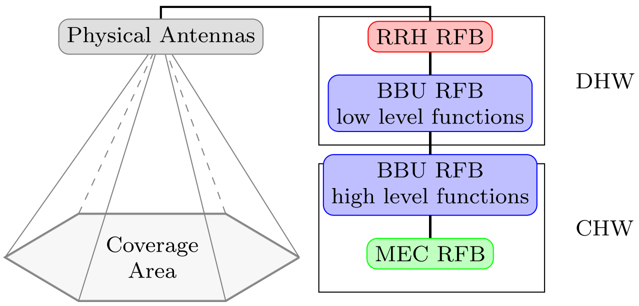


Fig. 1 Scheme of an RFB-based 5G node serving an area

a 5G node with one RRH RFB, one BBU RFB, and one MEC RFB. In general, each 5G node can pool also BBU RFBs and MEC RFBs from other nodes, e.g., by adopting a Cloud Radio Access Network (CRAN) paradigm [20]. As a result, the RFB chain is not constrained to be located on the same 5G node, but it can be realized across several nodes. Focusing on the resources consumed by the RFB on the HW, the RRH RFB and the BBU RFB consume an amount of bandwidth on the DHW of the node. In addition, we assume that the BBU RFB and the MEC RFB consume CPU and RAM resources on the CHW part of the node. The requirements in terms of consumed resources by the RFBs are then used in this work to properly dimension the 5G nodes.

Finally, we consider a further classification of each RFB task, which is based on the type. More in detail, Type 1 (T1) RFBs are used to serve large set of users. For example, a T1-RRH RFB can act as a macro cell, covering a vast portion of territory. On the other hand, T2 RFBs are instead used to serve small set of users. In this case, a T2-RRH RFB realizes a small cell. Clearly, the different RFB types are characterized by different requirements (in terms of bandwidth, CPU, and RAM) on the CHW and the DHW equipment. Given this taxonomy, we then detail in the following section how to minimize the total installation costs of an RFB-based network.

4 Optimal formulation

Let us denote with U and N the set of users and the set of 5G nodes, respectively. We then introduce the binary variable $x_{un} \in \{0, 1\}$, which takes value 1 if user $u \in U$ is served by an RRH RFB placed at node n, 0 otherwise. Each user is served by at most one node, which is expressed as:

$$\sum_{n \in N} x_{un} \leq 1 \qquad u \in U \tag{1}$$

Moreover, we introduce one single constraint to model that a minimum number of users has to be served:

$$\sum_{u \in U} \sum_{n \in N} x_{un} \geq \lceil \delta \cdot |U| \rceil \tag{2}$$

In this constraint, $\delta \in (0, 1]$ represents the minimum fraction of users that has to be covered by the 5G service, whereas $\lceil \cdot \rceil$ and $| \cdot |$ denote the ceiling of a number and the cardinality of a set, respectively.

In the following, we consider the installation constraints for the RRH RFBs. More in depth, we introduce the set R of RRH RFB types, and the binary variable y_{nr}^{RRH}, which takes value 1 if the RRH RFB of type $r \in R$ is installed at 5G node $n \in N$, 0 otherwise. Clearly, at most one type of RRH RFB can be installed in each node, so we have:

$$\sum_{r \in R} y_{nr}^{RRH} \leq 1 \qquad n \in N \tag{3}$$

In addition, we impose the fact that, if the node is serving a user, an RRH RFB has to be installed on it:

$$x_{un} \leq \sum_{r \in R(u)} y_{nr}^{RRH} \qquad u \in U, n \in N \tag{4}$$

where $R(u)$ denotes the subset of RRH RFB types that are compatible with a user $u \in U$.

The number of users served by each RRH RFB is then bounded by the maximum number of users that can be supported by the RRH RFB, which we denote as U_r^{max}. We express this condition with the following constraint:

$$\sum_{u \in U} x_{un} \leq \sum_{r \in R} U_r^{max} \; y_{nr}^{RRH} \qquad n \in N \tag{5}$$

In addition, we introduce the input parameters a_r^{RRH} to denote the number of available RRH RFBs of type $r \in R$. The total number of installed RRH RFBs must be less or equal than the available ones:

$$\sum_{n \in N} y_{nr}^{RRH} \leq a_r^{RRH} \qquad r \in R \tag{6}$$

We then consider the constraints relative to the BBU RFB and MEC RFB placement. In particular, we introduce the set B and the set M to store the BBU RFB types and the MEC RFB ones, respectively. We then denote with $v_{n_1 n_2 b}^{BBU}$ a binary variable taking the value of 1 if a BBU of type $b \in B$ placed at node $n_1 \in N$ serves the RFB chain originating from the RRH RFB placed at node $n_2 \in N$, 0 otherwise. Moreover, a_b^{BBU} is an input parameter, which stores the number of available BBU RFBs of type $b \in B$. The number of installed BBU RFBs is then bounded by a_b^{BBU} through the following constraint:

$$\sum_{n_1 \in N} \sum_{n_2 \in N} v_{n_1 n_2 b}^{BBU} \leq a_b^{BBU} \qquad b \in B \tag{7}$$

In a similar way, we limit the maximum number of used MEC RFBs through the following constraint:

$$\sum_{n_1 \in N} \sum_{n_2 \in N} v^{MEC}_{n_1 n_2 m} \leq a^{MEC}_m \qquad m \in M \tag{8}$$

where $v^{MEC}_{n_1 n_2 m}$ is a binary variable taking the value 1 if a MEC RFBs of type $m \in M$ is installed at node $n_1 \in N$ to serve the RFB chain originating from the RRH RFB placed at node $n_2 \in N$, 0 otherwise, and a^{MEC}_m is an input parameter storing the number of available MEC RFBs of type $m \in M$.

We then introduce the compatibility constraints between the RFBs. In particular, a BBU RFB can be part of the chain serving the RRH RFB placed in node $n_2 \in N$ only if it is compatible with that RRH RFB. We express this condition through the following constraint:

$$y^{RRH}_{n_2 r} \leq \sum_{n_1 \in N} \sum_{b \in B(r)} v^{BBU}_{n_1 n_2 b} \qquad n_2 \in N, r \in R \tag{9}$$

where $B(r)$ denotes the subset of BBU RFBs compatible with an RRH RFB of type $r \in R$. In a similar way, we introduce the compatibility constraint for the MEC RFBs:

$$y^{RRH}_{n_2 r} \leq \sum_{n_1 \in N} \sum_{m \in M(r)} v^{MEC}_{n_1 n_2 m} \qquad n_2 \in N, r \in R \tag{10}$$

where $M(r)$ is the subset of MEC RFBs that are compatible with an RRH RFB of type $r \in R$.

In the following, we consider the constraints governing the traffic from users. We then introduce the continuous variable $t_u \geq 0$ to store the amount of downlink traffic served to user $u \in U$. In addition, we introduce the input parameter CAP_{run}, which denotes the radio link capacity when user u is served by an RRH RFB of type r placed at node n. The amount of downlink traffic is then limited by the maximum radio link capacity:

$$t_u \; x_{un} \leq \sum_{r \in R} \text{CAP}_{run} \; y^{RRH}_{nr} \qquad u \in U, n \in N \tag{11}$$

The previous constraints are non-linear, since they contain the product of variables t_u and x_{un}. Such product can be linearized in a standard way (see, e.g., [21]) by introducing one *continuous* variable $\phi_{un} = t_u \; x_{un}$ and the four linear inequalities:

$$\phi_{un} \geq 0 \tag{12a}$$

$$\phi_{un} \leq \text{CAP}^{\max}_u \; x_{un} \tag{12b}$$

$$\phi_{un} \leq t_u \tag{12c}$$

$$\phi_{un} \geq t_u - (1 - x_{un}) \; \text{CAP}^{\max}_u \tag{12d}$$

where we have introduced the coefficient $\text{CAP}^{\max}_u = \max_{r \in R, n \in N}\{\text{CAP}_{run}\}$, for each $u \in U$. This substitution is correct since:

- if $x_{un} = 0$, then Eqs. 12a and 12b imply $\phi_{un} = 0$; additionally, Eq. 12c becomes $0 \leq t_u$ and Eq. 12d becomes $0 \geq t_u - \text{CAP}^{\max}_u$, which are both satisfied recalling that $0 \leq t_u \leq \text{CAP}^{\max}_u$ for each u;
- if $x_{un} = 1$, Eqs. 12c and 12d jointly give $\phi_{un} = t_u$ and Eqs. 12a and 12b provide the (correct) bounds $0 \leq \phi_{un} \leq \text{CAP}^{\max}_u$.

The linear version of constraint (11) is then:

$$\phi_{un} \leq \sum_{r \in R} \text{CAP}_{run} \; y^{RRH}_{nr} \qquad u \in U, n \in N \tag{13}$$

Moreover, the total capacity provided to the connected users has to be lower than the maximum total capacity managed by an RRH RFB of type r, which we denote as CAP^{RRH}_r. We express this condition with the following constraint:

$$\sum_{u \in U} \text{CAP}_{run} x_{un} \; y^{RRH}_{nr} \leq \text{CAP}^{RRH}_r \qquad n \in N, r \in R \tag{14}$$

Similarly to constraint (11), we linearize the product $x_{un} \; y^{RRH}_{nr}$ by introducing a new continuous variable $\theta_{unr} = x_{un} \; y^{RRH}_{nr}$ accompanied by the four constraints:

$$\theta_{unr} \geq 0 \tag{15a}$$

$$\theta_{unr} \leq x_{un} \tag{15b}$$

$$\theta_{unr} \leq y^{RRH}_{nr} \tag{15c}$$

$$\theta_{unr} \geq x_{un} + y^{RRH}_{nr} - 1 \tag{15d}$$

The linear version of constraint (14) is then:

$$\sum_{u \in U} \text{CAP}_{run} \theta_{unr} \leq \text{CAP}^{RRH}_r \qquad n \in N, r \in R \tag{16}$$

We then introduce the input parameter CAP^{MEC}_m, which is used to denote the maximum capacity that can be managed by a MEC RFB of type m. The total traffic from users connected to the RRH RFB placed at node n_1 has to be lower than the maximum capacity managed by the MEC RFB in the chain:

$$\sum_{u \in U} \sum_{n_1 \in N} t_u x_{un_1} v^{MEC}_{n_1 n_2 m} \leq \text{CAP}^{MEC}_m \sum_{n_1 \in N} v^{MEC}_{n_1 n_2 m}, \qquad n_2 \in N, m \in M \tag{17}$$

Also in this case, we face a non-linear constraint containing the product of (three) variables. To linearize it, similarly to what we have done for Eq. 11, we first use the linearization variables introduced in Eq. 13, imposing $\phi_{un_1} = t_u x_{un_1}$; then we face the resulting product of variables $\phi_{un_1} v^{MEC}_{n_1 n_2 m}$, which can be linearized by introducing a new continuous variable $\varphi_{un_1n_2m} = \phi_{un_1} v^{MEC}_{n_1n_2m}$ and the following four constraints:

$$\varphi_{un_1n_2m} \geq 0 \tag{18a}$$

$$\varphi_{un_1n_2m} \leq \text{CAP}^{\max}_u \; v^{MEC}_{n_1n_2m} \tag{18b}$$

$$\varphi_{un_1n_2m} \leq \phi_{un_1} \tag{18c}$$

$$\varphi_{un_1n_2m} \geq \phi_{un_1} - (1 - v^{MEC}_{n_1n_2m}) \; \text{CAP}^{\max}_u \tag{18d}$$

The linear version of constraint (17) is then:

$$\sum_{u \in U} \sum_{n_1 \in N} \varphi_{un_1n_2m} \leq \text{CAP}^{MEC}_m \sum_{n_1 \in N} v^{MEC}_{n_1n_2m}, \qquad n_2 \in N, m \in M \tag{19}$$

Moreover, as input to the problem, we introduce a set $CONF_r$ that includes all the pairs of nodes that conflict for an RRH RFB type $r \in R$: if a pair (n_1, n_2) belongs to $CONF_r$, then at most one RRH RFB of type r can be installed either in n_1 or in n_2. Formally, this is expressed by the constraint:

$$y^{RRH}_{n_1r} + y^{RRH}_{n_2r} \leq 1 \qquad r \in R, (n_1, n_2) \in CONF_r \tag{20}$$

In addition, we impose the fact that the MEC RFBs and the BBU RFBs can be installed only in nodes already storing RRH RFBs:

$$v^{MEC}_{n_1n_2m} \leq y^{RRH}_{n_1r} \qquad r \in R, n_1, n_2 \in N, m \in M \tag{21}$$

$$v^{BBU}_{n_1n_2b} \leq y^{RRH}_{n_1r} \qquad r \in R, n_1, n_2 \in N, b \in M \tag{22}$$

In the following, we impose that the traffic assigned to users has to be higher than a minimum value, denoted with t^{MIN}:

$$t_u \geq t^{MIN} x_{un} \qquad u \in U, n \in N \tag{23}$$

Finally, we consider the CAPEX costs. Let us denote with c^{SITE}_r the cost for installing a site able to host an RRH RFB of type r. In addition, we denote with c^{CH} and c^{DH} the costs for installing the CHW and the DHW at the node, respectively. Moreover, let us denote with c^{BBU}_b and c^{MEC}_m the costs for installing one BBU RFB of type b and one MEC RFB of type m, respectively.

The OPTIMAL 5G DESIGN (OPT-5GD) is then defined as:

$$\min \sum_{n \in N} \sum_{r \in R} \left(c^{SITE}_r + c^{CH} + c^{DH} \right) y^{RRH}_{rn} + \\ + \sum_{n_1 \in N} \sum_{n_2 \in N} \left(\sum_{b \in B} c^{BBU}_b v^{BBU}_{n_1n_2b} + \sum_{m \in M} c^{MEC}_m v^{MEC}_{n_1n_2m} \right) \tag{24}$$

Users to RRH RFB assignment:	Eq. (1), (2)
RRH RFB installation constraints:	Eq. (3), (4)
Maximum number of users per RRH RFB	Eq. (5)
Maximum number of available RFBs	Eq.(6), (7), (8)
RFB chain compatibility constraints	Eq. (9), (10)
Maximum RRH RFB capacity	Eq. (11), (12)
Maximum MEC RFB capacity	Eq. (13)
RRH RFB conflict constraint	Eq. (14)
MEC/BBU RFB placement constraints	Eq. (15), (16)
Minimum traffic constraints	Eq. (17)
Linearization constraints	

Eq.$(12a - 12d)$, $(15a - 15d)$, $(18a - 18d)$

(25)

Under variables: $x_{un} \in \{0, 1\}$, $t_u \geq 0$, $y^{RRH}_{nr} \in \{0, 1\}$, $v^{BBU}_{n_1n_2b} \in \{0, 1\}$, $v^{MEC}_{n_1n_2m} \in \{0, 1\}$, $\phi_{un_1} \geq 0$, $\theta_{unr} \geq 0$, $\varphi_{un_1n_2m} \geq 0$.

Proposition 1 *The* OPT-5GD *problem is NP-Hard.*

Proof In order to prove the statement, we show that a subproblem of OPT-5GD obtained by keeping only the decision variables $v^{MEC}_{n_1n_2m} \in \{0, 1\}$ $\forall n_1, n_2 \in N, m \in M$ and fixing all the remaining decision variables x_{un} $\forall u \in U, n \in N$, y^{RRH}_{nr} $\forall n \in N, r \in R$, $v^{BBU}_{n_1n_2b}$ $\forall n_1, n_2 \in N, b \in B$, t_u $\forall u \in U$ to a feasible combination of values $\bar{x}_{un}$, $\bar{y}^{RRH}_{nr}$, $\bar{v}^{BBU}_{n_1n_2b}$, $\bar{t}_u$ leads to an NP-Hard problem.

Moreover, in this subproblem we also consider the special case where only one type of RRH, BBU, and MEC RFBs are available (i.e., $|R| = |B| = |M| = 1$) and we can thus drop the indices r, b, m in all constraints and parameters. Under this setting, it is easy to check that the subproblem that we face thus reduces to:

$$\min \sum_{n_1 \in N} \sum_{n_2 \in N} c^{MEC} v^{MEC}_{n_1n_2} \tag{26}$$

$$\sum_{n_1 \in N} \sum_{n_2 \in N} v^{MEC}_{n_1n_2} \leq a^{MEC} \tag{27}$$

$$\bar{y}^{RRH}_{n_2r} \leq \sum_{n_1 \in N} v^{MEC}_{n_1n_2} \quad n_2 \in N \tag{28}$$

$$\sum_{u \in U} \sum_{n_1 \in N} \bar{t}_u \bar{x}_{un_1} v^{MEC}_{n_1n_2} \leq \text{CAP}^{MEC} \sum_{n_1 \in N} v^{MEC}_{n_1n_2}$$

$$n_2 \in N \tag{29}$$

$$v^{MEC}_{n_1 n_2} \in \{0, 1\} \quad n_1, n_2 \in N \tag{30}$$

We remark that in this subproblem the variables and constraints introduced to replace the product of decision variables are not needed. This problem is actually a generalization of the well-known *multiple knapsack problem* that additionally includes multiple knapsack constraints (29) and cardinality constraints imposing upper (27) and lower bounds (28) on the activation of decision variables representing putting items in the knapsacks. Such generalization is NP-Hard (see, e.g., [22, 23]) and thus also the complete problem OPT-5GD that we face is NP-Hard. □

Algorithm 1 Pseudocode of the SuperFluid Design Algorithm (SFDA).

```
Input: $N$, $U$, $a_r^{RRH}$, $a_b^{BBU}$, $a_m^{MEC}$, $CAP_{run}$, $t^{MIN}$, $\delta$, order_type
Output: $y_{nr}^{RRH}$, $v_{n_1 n_2 b}^{BBU}$, $v_{n_1 n_2 b}^{MEC}$, $x_{un}$
tot_cost_best_conf=Inf;
all_conf=comp_conf($N$, $a_r^{RRH}$, $r = 1$);
for curr_conf in all_conf do
    tot_RRH_RFB=0;
    u_cand_served= comp_cand_served_u(curr_conf, order_type, $U$, $t^{MIN}$);
    n_sorted=sort_RRH_RFB(u_cand_served, curr_conf, $r = 1$);
    curr_u_to_serve=$U$;
    for $n$ in n_sorted do
        u_assoc=associate_u($n$, curr_u_to_serve, $t^{MIN}$, $r = 1$);
        curr_u_to_serve=remove_served_u($U$, u_assoc);
    end for
    n_sorted=sort_RRH_RFB(curr_u_to_serve, $N$, $r = 2$)
    for $n$ in n_sorted do
        if check_tot_u_served(u_assoc,$\delta$)==false) then
            if (check_conf(curr_conf,$n$, $r = 2$)==true)&& (tot_RRH_RFB< $a_{r=2}^{RRH}$) then
                tot_RRH_RFB=tot_RRH_RFB+1;
                curr_conf=add_RRH_RFB(curr_conf, $n$, $r = 2$);
                u_assoc=associate_u($n$, curr_u_to_serve, $t^{MIN}$, $r = 2$);
                curr_u_to_serve=remove_served_u($U$, u_assoc);
            end if
        end if
    end for
    curr_conf=add_BBU_MEC_RFB(curr_conf, u_assoc, $a_b^{BBU}$, $a_m^{MEC}$, $t^{MIN}$);
    tot_cost=comp_tot_cost(curr_conf);
    if (tot_cost<tot_cost_best_conf)&& (check_tot_u_served(u_assoc,$\delta$)==true) then
        tot_cost_best_conf=tot_cost;
        [$y_{nr}^{RRH}$, $v_{n_1 n_2 b}^{BBU}$, $v_{n_1 n_2 b}^{MEC}$, $x_{un}$]= save_conf(curr_conf, u_assoc, $t^{MIN}$);
    end if
end for
```

Since the aforementioned formulation may be challenging to be solved in a realistic scenario, we propose in the next section two efficient algorithms to solve it.

5 Description of the algorithms

We initially describe the SuperFluid Design Algorithm (SFDA), then we detail the 5G Performance Clustered Design Algorithm (5G-PCDA), and finally we discuss the computational complexity of the two heuristics.

5.1 SuperFluid design algorithm

We design the SFDA algorithm by adopting a *divide et impera* approach, in which first the T1-RRH RFBs are placed and then the T2-RRH RFBs are installed. Then, once the RRH RFBs are placed, the algorithm performs the assignment of the MEC RFBs and the BBU RFBs. The goal of SFDA is to reduce as much as possible the CAPEX costs, while ensuring an adequate Quality of Service (QoS) to users. The main intuitions behind this approach are the following ones: (i) the T1-RRH RFBs are actually acting as macro cells; their number is lower compared with T2-RRH RFBs, which are instead used as small cells, (ii) the main goal of the T1-RRH RFBs is to provide coverage over the territory, and to guarantee the service to the largest number of users, (iii) T2-RRH RFBs are used to provide capacity to a subset of users, i.e., the ones falling in their coverage area, which is clearly lower than the coverage area of T2-RRH RFBs, and (iv) once the RRH RFBs are placed, the installation of the BBU RFBs and MEC RFBs is performed considering the same subset of nodes hosting the RRH RFBs.

Alg. 1 reports the pseudo-code of the proposed solution. The algorithm requires as input the set of candidate nodes N, the set of users U, the numbers of available RFBs a_r^{RRH}, a_b^{BBU}, a_m^{MEC} (for each type), the downlink capacity CAP_{run}, the threshold δ, and the traffic per user t^{MIN}. In addition, a sorting rule, denoted as order_type in Alg. 1, is required for the ordering of the T1-RRH RFBs. More in detail, we consider the following ordering criteria: (i) descending number of users that can be served by each T1-RRH RFB or (ii) descending number of users that can be served by each T1-RRH RFB but cannot be served by any T2-RRH RFBs. The rationale behind these criteria is the following: the first one aims to cover as much users as possible, while the second is restricted to serve users that can not be served by any T2-RRH RFBs, due, e.g., to large distance and/or the presence of obstacles between the user and the cell. In other words, such users would be not served at all by any RRH RFB, unless a proper configuration of T1-RRH RFBs is installed. The actual choice between the two criteria is left as input parameter to SFDA.

Initially, the total cost for the best configuration is initialized to a very large value (line 3). Moreover, the algorithm computes all the possible configurations for placing the T1-RRH RFBs over the considered scenario (line 4). More in detail, the actual number of nodes that can host the T1-RRH RFBs is normally pretty limited, due to multiple reasons: (i) the number of available T1-RRH RFBs is limited, (ii) T1-RRH RFBs should be placed not so close to each other (to limit the impact of interference), and (iii) users living in the scenario are not willing that the operator installs a large number of T1-RRH RFBs over them. Then, for each possible configuration of T1-RRH RFBs (line 5), the algorithm initially computes the users that can be served by the current configuration in terms of installed T1-RRH RFBs (line 7). In the following, the T1-RRH RFBs are ordered (line 8), based on one of the aforementioned sorting criteria. The current set of users to serve is then initialized to the total number of users (line 9). Finally, for each T1-RRH RFB, the users are associated to the current cell (line 10), and the current set of users that need to be served is updated (lines 11–12).

In the following step, the T2-RRH RFBs are sorted, based on the number of users that can be served by each of them (line 14). For each T2-RRH RFB (line 15), if there are still users to be served (line 16), a check on the current configuration is performed (line 17). In particular, the current T2-RRH RFB can be installed on node n only if: (i) n is not in conflict with the current configuration (e.g., the current node n is not already in use by a T1-RRH RFB, and/or a minimum distance between the RRH RFBs of the same type is ensured), and (ii) the number of used T2-RRH RFBs is lower than the available one. If both conditions hold, the total number of used T2-RRH RFBs is incremented (line 18), the current configuration is updated (line 19), and both the users that are associated and the ones that need to be served are updated (lines 20–21).

Once the RRH RFBs are placed, the MEC RFBs and the BBU RFBs are installed (line 25). The rule to install these RFBs is straightforward: the same type of MEC RFB and BBU RFB is installed on each node hosting a given type of RRH RFB. In other words, the entire RFB chain for an RRH RFB is located on the same node hosting the RRH RFB. Moreover, the total cost of the current configuration is computed (line 26), and the best cost, as well as the best configuration, are eventually updated (lines 27–30). At the end of the procedure, SFDA produces as ouput the set of installed RFBs, as well as the assignment of each user to each RRH RFB.

5.2 5G performance clustered design algorithm

We then detail the 5G Performance Clustered Design Algorithm (5G-PCDA). The goal of 5G-PCDA is to increase as much as possible the number of served users, by targeting also the reduction in the algorithm complexity. Alg. 2 reports the pseudo-code of 5G-PCDA. Clearly, the same input (except from δ) and the same output of SFDA are required. Initially, 5G-PCDA computes the number of users that can be potentially served by placing the T1-RRH RFB in each candidate site (line 3). The descending number of users that can be served by each T1-RRH RFB is taken as ordering rule. Then, the RRH RFBs which can be potentially installed are sorted (line 4). In the following,

Algorithm 2 Pseudocode of the 5G Performance Clustered Design Algorithm (5G-PCDA).

```
1: Input: $N, U, a_r^{RRH}, a_b^{BBU}, a_m^{MEC}, CAP_{run}, t^{MIN}$, order_type
2: Output: $y_{nr}^{RRH}, v_{n_1 n_2 b}^{BBU}, v_{n_1 n_2 b}^{MEC}, x_{un}$
3: u_cand_served= comp_cand_served_u(curr_conf, order_type, $U, t^{MIN}$);
4: RRH_sorted=sort_RRH_RFB(u_cand_served, curr_conf, $r = 1$);
5: n_users_prec_conf=0
6: users_to_serve=$U$;
7: for i=1; i $\leq a_1^{RRH}$; i++ do
8:     if check_conf(curr_conf, RRH_sorted[i], $r = 1$)==true) then
9:         curr_conf=install_RRH(RRH_sorted[i], $r = 1$);
10:        [u_assoc users_to_serve]=associate_u(curr_conf, users_to_serve, $t^{MIN}$, $r = 1$);
11:        if size(u_assoc) == n_users_prec_conf then
12:            curr_conf=uninstall_RRH(RRH_sorted[i]);
13:        else
14:            n_users_prec_conf=size(u_assoc);
15:        end if
16:    end if
17: end for
18: RRH_density=comp_dens_RRH(users_to_serve, curr_conf);
19: RRH_sorted=sort_RRH_RFB(RRH_density, curr_conf, $r = 2$);
20: for i=1; i $\leq a_2^{RRH}$; i++ do
21:     if check_conf(curr_conf, RRH_sorted[i], $r = 2$)==true) then
22:         curr_conf=install_RRH(RRH_sorted[i], $r = 2$);
23:         u_assoc users_to_serve]=associate_u(curr_conf, users_to_serve, $t^{MIN}$, $r = 2$);
24:         if size(u_assoc) == n_users_prec_conf then
25:             curr_conf=uninstall_RRH(RRH_sorted[i]);
26:         else
27:             n_users_prec_conf=size(u_assoc);
28:         end if
29:     end if
30: end for
31: curr_conf=add_BBU_MEC_RFB(curr_conf, u_assoc, $a_b^{BBU}$, $a_m^{MEC}$, $t^{MIN}$);
32: [$y_{nr}^{RRH}$, $v_{n_1 n_2 b}^{BBU}$, $v_{n_1 n_2 b}^{MEC}$, $x_{un}$]= save_conf(curr_conf, u_assoc, $t^{MIN}$);
```

the algorithm iteratively installs each T1-RRH RFBs (line 7-17). More in detail, the current T1-RRH RFB is installed only if it is compatible with the current configuration (lines 8–9). In the following, the association of the users with the current set of installed RRH RFBs is computed (line 10). If the number of associated users of the current configuration is the same as the number of associated users in the previous iteration, the algorithm greedily decides to not install the current RRH RFB (line 11–12). Otherwise, the current RRH RFB is kept, and the number of users served by the current configuration is stored (line 14). In the second part of the algorithm (lines 18–30), the T2-RRH RFBs are installed in order to serve the remaining users. Firstly, a grid of regular square size is applied to the territory under consideration. For each cell in the grid, the cell density is computed as the number of users falling inside the current cell. In the following, we associate to each candidate T2-RRH RFB the density value of the cell that includes the position of the current RFB. Both the two steps are performed in the `comp_dens_RRH` function of line 18. Given the values of RRH density, our goal is then to select the candidate T2-RRH RFBs potentially able to serve the highest number of users. To do that, we sort the T2-RRH RFBs by decreasing RFB density (line 19). The algorithm then try to iteratively install the T2-RRH RFBs (line 20–30). For each candidate T2-RRH RFB, a check about the compatibility with the current configuration is performed (line 21). If the current RRH RFB is compatible with the current configuration, the association of the users to the RRH RFB is performed (line 23). If there is not an improvement in the number of associated users, the current RRH RFB is uninstalled (lines 24–25), otherwise it is kept, and the current number of served users is updated (line 27). Finally, in the last steps of the algorithm, the BBU and MEC RFB are installed (line 31), and the resulting configuration is saved (line 32), considering the same functions used by SFDA.

5.3 Computational complexity

We then evaluate the computational complexity of the proposed algorithms. Focusing on SFDA, the computation of all the possible configurations in line 4 of Alg. 1 results in $\mathcal{O}(|N|!)$. Focusing then on the computation of the number of users that can be served by each T1-RRH RFB (line 7), its complexity is in the order of $\mathcal{O}(|N| \times |U|))$. The sorting of the T1-RRH RFBs (line 8) has a complexity of $\mathcal{O}(|N| \times \log(|N|))$. The association of the users to the installed RRH RFBs (lines 10–13) has a complexity of $\mathcal{O}(|N| \times |U|)$. The sorting of the T2-RRH RFBs has a complexity of $\mathcal{O}(|N| \times \log(|N|))$. Similarly to the T1-case, also the association of users to the T2-RRH RFBs has a complexity of $\mathcal{O}(|N| \times |U|)$. Moreover, the association of the MEC/BBU RFBs requires $\mathcal{O}(|N| \times |U|)$. Finally, the saving of the best configuration results in a complexity of $\mathcal{O}(|N|^2 \times |B| + |N| \times |U| + |N| \times |R|)$. Overall, the complexity of SFDA is in the order of $\mathcal{O}(|N|! \times (|N| \times \log(|N|) + |N| \times |U| + |N|^2 \times |B| + |N| \times |R|))$.

Focusing on 5G-PCDA, the computation of the number of users that can be potentially served (line 3 of Alg. 2) and the sorting of the candidate sites (line 4) have a complexity of $\mathcal{O}(|N| \times |U|))$ and $\mathcal{O}(|N| \times \log(|N|)$, respectively. The check of the current configuration (line 8) and the installation of the current RFB (line 9) have a complexity of $\mathcal{O}(|N|)$. Then, the association of users to the current configuration (line 10) has a complexity of $\mathcal{O}(|N| \times |U|)$. Clearly, the deallocation of the current RFB has a complexity of $\mathcal{O}(|N|)$ (line 12). The entire cycle over the T1-RRH RFBs (lines 7–17) has a complexity of $\mathcal{O}(|N|^2 \times |U|)$. In addition, the installation of the T2-RRH RFBs (lines 18–30) requires an array of cells to perform the grid density computation. Let us denote with $|G|$ the required number of cells in the grid. Clearly, the association of each user to a cell has a complexity of $\mathcal{O}(|U| \times |G|)$. In the following, the computation of the T2-RRH RFB density is done in $\mathcal{O}(|N| \times |G|)$. The remaining steps have then the same complexity as the T1-RRH RFB installation. Finally, the complexity of the last two steps (lines 31–32) are $\mathcal{O}(|N| \times |U|)$ and $\mathcal{O}(|N|^2 \times |B| + |N| \times |U| + |N| \times |R|)$, respectively. Overall, the complexity of 5G-PCDA is $\mathcal{O}(|N|^2 \times |U| + |N| \log(|N|) + |N|^2 \times |B| + |N| \times |R| + |U| \times |G| + |N| \times |G|)$.

6 Scenarios and parameters settings

The data we use is based on real-world dataset coming from the WeFi app [6], processed as described in [24]. The WeFi dataset was collected in October 2015 in a 11×11 km^2 area corresponding to the city and county of San Francisco. The dataset is a collection of over nine million *records*, each of them containing:

- day, hour (a coarse-grained timestamp);
- anonymized user identifier and GPS position;
- Mobile Network Operator (MNO), cell ID, cell technology (e.g., 3G/4G);
- Wi-Fi network (SSID) and access point (BSSID) the user is connected to (if any);
- active app and amount of downloaded/uploaded data.

If the position of the user or the networks he/she is connected change within a 1-h period, multiple records are generated. Similarly, one record is generated for each app that is active during the same period. Overall, the dataset contains information about 7182 unique users and 78,948 cell IDs. Unlike similar datasets that are provided by mobile operators, the WeFi one is *crowd-sourced*, i.e., contributed directly by users of the WeFi app. Its crowd-sourced nature allows it to include information about multiple network technologies (e.g., cellular and Wi-Fi) as well as multiple mobile operators. Due to licensing issues, the WeFi dataset cannot directly be employed in research. As described in [24], it is instead leveraged to train a set of distributions which are in turn used to obtain a new trace, distinct from the WeFi one but exhibiting the same space- and time-related features, e.g., data demand patterns and infrastructure deployment. To give more insights, Fig. 2 reports the positions of the sites and the users over the territory.

Over the whole dataset, we select two representative scenarios, namely (i) a portion of 1000 × 1000 [m^2] of the city center and henceforth named "SAN Small" and (ii) a portion of 3600 × 3700 [m^2] including the downtown area and henceforth named "SAN Big". In addition, the provided positions of the candidate sites are used for placing T1-RRH RFBs. On the other hand, we consider as candidate sites to install the T2-RRH RFBs the points at the interesections of a square grid, with a distance of 100 [m] between any two consecutive points. Figure 3 reports the positions of the candidate sites and the users over the considered scenarios.

Given the two scenarios, we set the input parameters, which are summarized in Table 1. Unless otherwise specified, we adopt a similar setting of input parameters as in [7]. More in detail, the T1-RRH RFB is able to serve more users compared with the T2-RRH RFB. In addition, we consider a relatively lower number of available T1 RFBs compared with T2 RFBs. Both numbers are set equal to the cardinality of the number of candidate sites in each scenario. Focusing then on the downlink capacity model, we adopt the same model of Marzetta [18]. We refer the reader to [7] for a detailed description of the parameters adopted for this model. Moreover, the compatibility matrix of possible configurations $CONF_r$ is set in accordance to the following rules: (i) each pair of T1-RRH RFB nodes has always to guarantee a minimum distance of 400 [m] between them, (ii) the minimum distance for placing T2-RRH RFBs is set equal to 50 [m]. In this way, we limit the negative effect of placing two T1-RRH RFBs too close to each other, while we allow the T2-RRH RFBs to be installed potentially

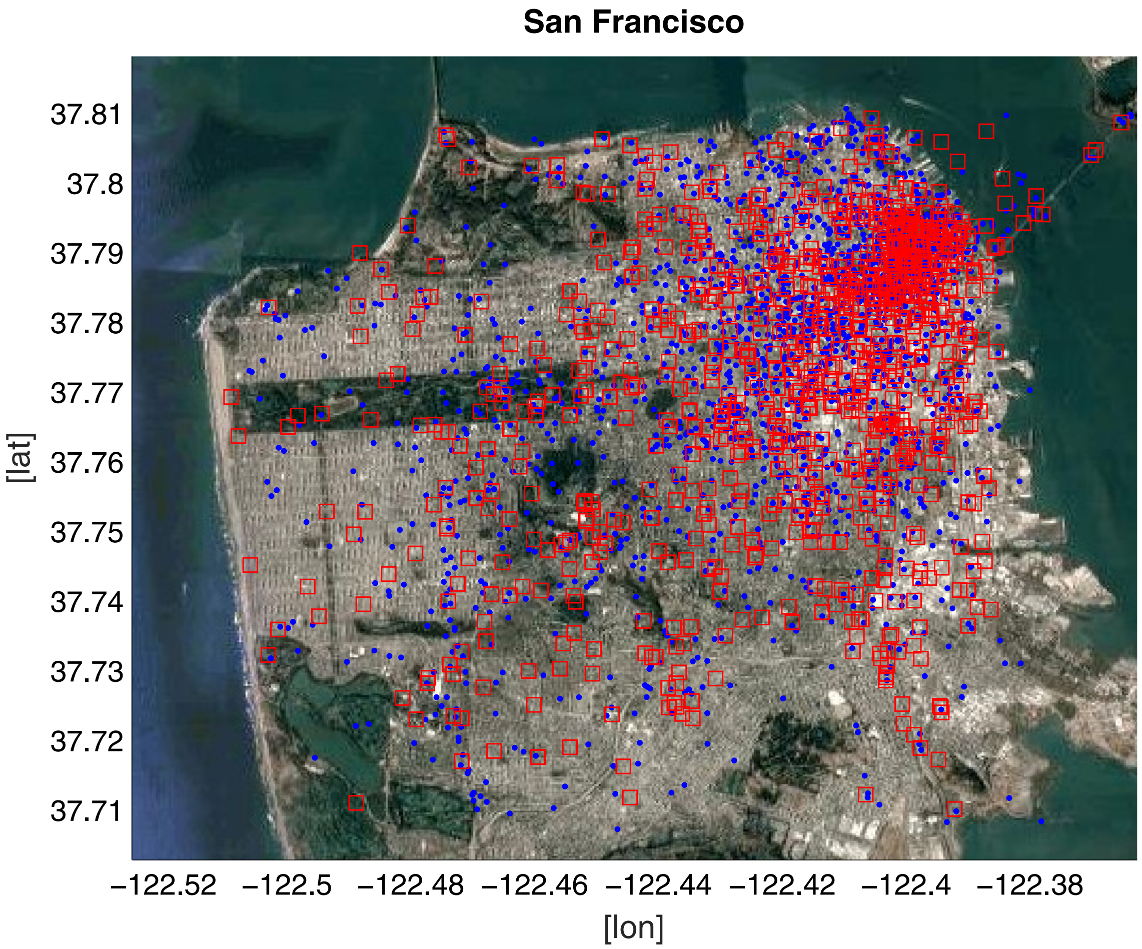
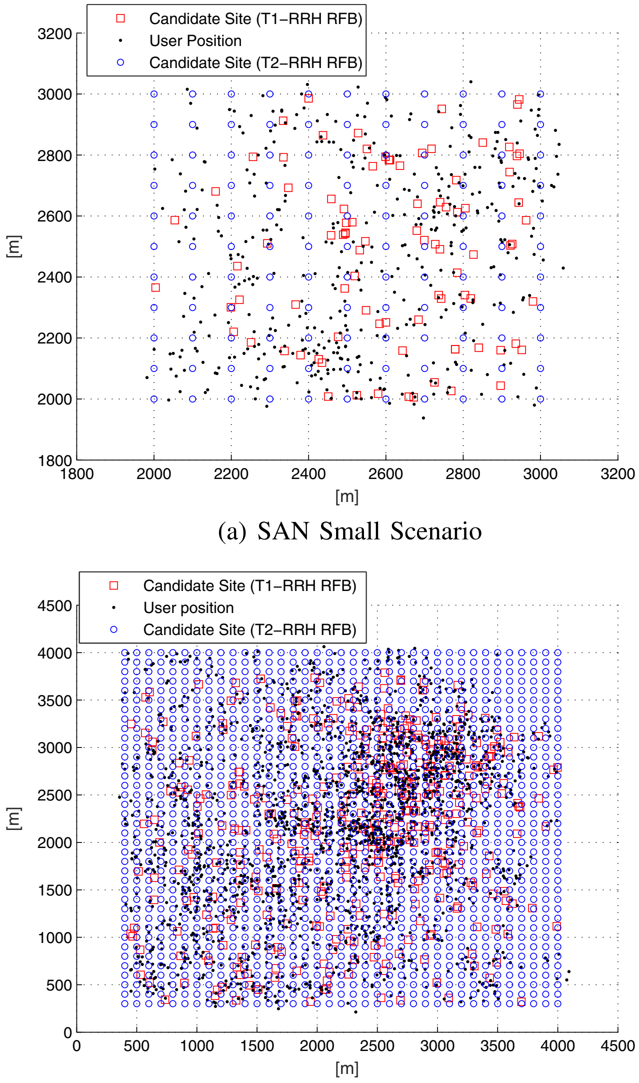


Fig. 2 San Francisco dataset: positions of the sites and the users from the WeFi app



(a) SAN Small Scenario

(b) SAN Big Scenario

Fig. 3 Candidate sites and user positions for the two considered scenarios

in each site. Focusing then on the costs, we assume that the site installation costs are higher for the nodes hosting T1-RRH RFBs compared with the ones running T2-RRH RFBs. Focusing on the CHW and DHW costs, we assume two distinct fixed terms, that have to be paid if the node is installed (independently from the RFB type), plus two additional terms that depends on the number and on the type of BBU RFBs and MEC RFBs installed on the node. The rationale behind this setting is the following one: actually, both BBU RFB and MEC RFB consume a large amount of Random Access Memory (RAM) [25], which has to be properly dimensioned. Note that in [25] we consider only two costs related to memory installation when a T1-RFB or a T2-RFB is installed. Here, instead, we consider a more general case, in which the costs depends on the number and types of installed RFBs.

7 Performance evaluation

We evaluate the SFDA and 5G-PCDA algorithms over the considered scenarios. Unless otherwise specified, we assume a grid of size 800 × 800 [m^2] for the 5G-PCDA algorithm. In addition, in order to introduce a term of comparison, we also code a classical first-fit algorithm [26], referred as First Fit Design Algorithm (FFDA). The main goal of FFDA is to greedily iterate over the set of users and the set of candidate T1- and T2-RRH RFBs. For each user and each candidate RRH RFB, a check on the current RFB is performed. In particular, if the current RFB can serve the user and it is already installed, then the user is associated to the current RRH RFB. Otherwise, if the current RFB can serve the user but it is not installed, a check on the compatibility with the already installed RRH RFB is performed. If it is possible to install the current RRH RFB, then the user is associated to it. Finally, the BBU and MEC RFBs are placed according to the same rule of SFDA and 5G-PCDA. Clearly, we expect that a large number of resources is installed by FFDA, due to the fact that this solution does not optimize the costs and the traffic requests from users. Finally, all the algorithms have been coded in Matlab, and they have been run on a laptop equipped with 2 cores Intel Core i7 at 2.8 [GHz] and 8 [GB] of RAM.

7.1 Results from SAN small scenario

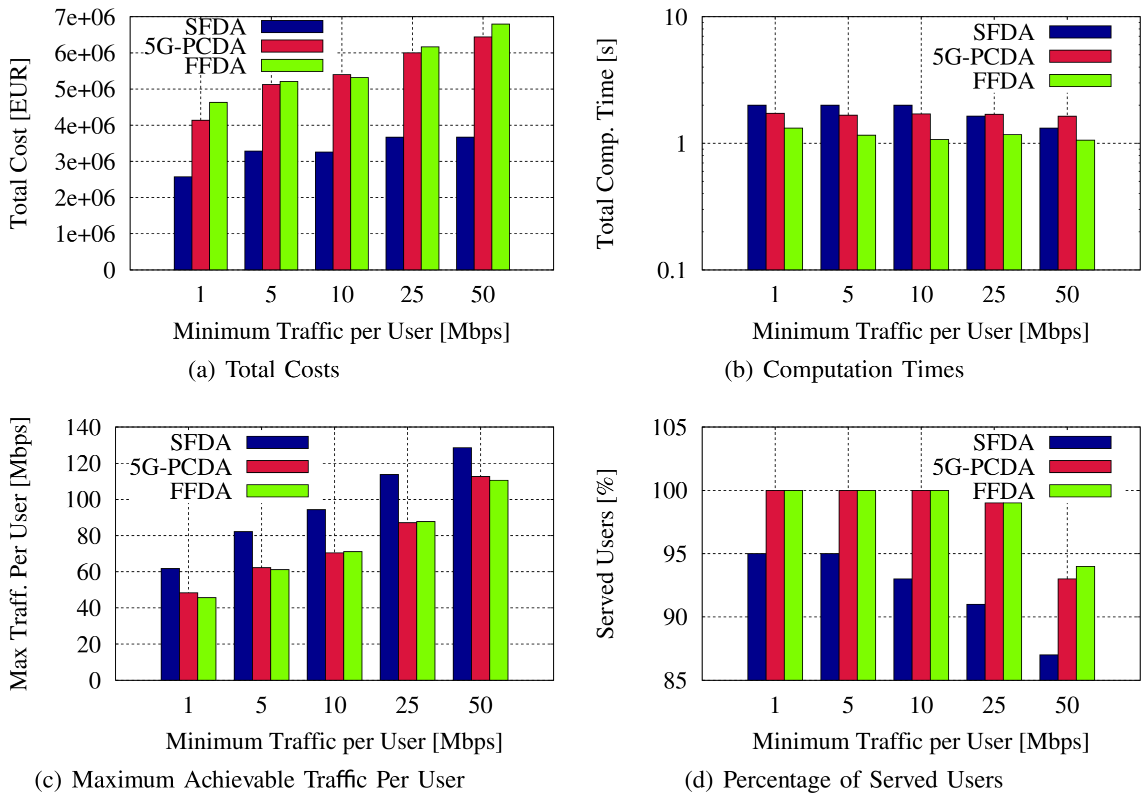
We initially evaluate the impact of varying the minimum amount of traffic t^{MIN} between 1 and 50 [Mbps]. Moreover, we set the δ threshold equal to 85% for SFDA. Figure 4a reports the total costs vs. the variation of t^{MIN}. As expected, the costs are increasing when t^{MIN} is increased, due to the fact that more RFBs and 5G nodes have to be installed in order to fulfill the traffic requirements. However, we can see that the costs experience an increase of less than two times when t^{MIN} passes from 1 to 50 [Mbps]. The relatively small increase of the total costs compared with the sharp increase of traffic is due to following reasons: (i) an amount of resources has to be installed in any case, in order to provide coverage to users (i.e., independently from their amount of requested traffic), (ii) when the resources are installed, it is possible to exploit their capacity in order to provide the requested service to users. In addition, we can note that 5G-PCDA requires an higher amount of additional costs compared with SFDA. This is an expected result, being the main goal of 5G-PCDA the maximization of the number of served users. Moreover, we can note that the performance of 5G-PCDA is similar to FFDA. This is also an expected

Table 1 Input parameters

Parameter	Value	
$\lvert U \rvert$	431 (SAN Small) - 1960 (SAN Big)	
$\lvert N \rvert$	212 (SAN Small) - 1832 (SAN Big)	
U_r^{MAX}	T1-RRH RFB: 126	T2-RRH RFB: 42
a_r^{RRH}	T1-RRH RFB: 91 (SAN Small) - 426 (SAN Big)	T2-RRH RFB: 121 (SAN Small) - 1406 (SAN Big)
a_b^{BBU}	T1-RRH RFB: 91 (SAN Small) - 426 (SAN Big)	T2-RRH RFB: 121 (SAN Small) - 1406 (SAN Big)
a_m^{MEC}	T1-RRH RFB: 91 (SAN Small) - 426 (SAN Big)	T2-RRH RFB: 121 (SAN Small) - 1406 (SAN Big)
CAP_{run}	Model from Marzetta [18] with input parameters from [7].	
CAP_r^{RRH}	T1-RRH RFB: 30 [Gbps]	T2-RRH RFB: 10 [Gbps]
CAP_m^{RRH}	30 [Gbps] (T1-MEC RFB, T2-MEC RFB)	
$CONF_r$	Compatibility matrix ensuring 400 [m] of minimum distance among T1-RRH RFBs and 50 [m] of minimum distance among T2-RRH RFBs.	
t^{MIN}	1-50 [Mbps]	
c_r^{SITE}	T1-RRH RFB: 120 [k€]	T2-RRH RFB: 40 [k€]
c^{CH}	4711 [€]	
c^{DW}	9240 [€]	
c_b^{BBU}	T1-BBU RFB: 1307 [€]	T2-BBU RFB: 440 [€]
c_m^{MEC}	T1-BBU RFB: 1307 [€]	T2-BBU RFB: 440 [€]

result, since, in this scenario, a large number of resources is installed by 5G-PCDA.

In the following, we consider the impact of computation times, as reported in Fig. 4b. Interestingly, all the algorithms experience a relatively low computation time, i.e., at most 2 [s]. This is due to the fact that the scenario is relatively small, and therefore the computation of all the possible set of candidate sites to host T1-RRH RFB done by SFDA is



(a) Total Costs

(b) Computation Times

(c) Maximum Achievable Traffic Per User

(d) Percentage of Served Users

Fig. 4 Performance of the SFDA, 5G-PCDA, and FFDA algorithms vs. the minimum traffic per user t^{MIN} over the SAN Small scenario

pretty feasible. Moreover, also the other steps of both the algorithms can be performed in few seconds.

We then consider the maximum amount of traffic that can be served to each user. In particular, given the output of both SFDA, 5G-PCDA, and FFDA in terms of assignment of users to the 5G nodes x_{un} and type of RRH RFB installed y_{nr}^{RRH}, we set $t_u = CAP_{run}$ for each user u, each node n, and each type r holding $x_{un} = 1$ and $y_{nr}^{RRH} = 1$. In this way, we compute the maximum amount of traffic that can be served to the users. Figure 4c reports the obtained results. Interestingly, all the solutions are able to provide a large throughput to users, i.e., more than 40 [Mbps], even when $t^{MIN} = 1$ [Mbps]. This is due to the fact that the capacity of the installed RFBs is able to ensure large requests of traffic from users. However, we point out that the actual amount of traffic served to each user (i.e., which may ba larger than t^{MIN}) is done during the management phase, in order to accomplish to possible traffic variations. We leave the investigation of this last aspect as future work. In any case, however, we can note that the maximum amount of traffic increases with increasing values of t^{MIN}, due to the fact that more resources, in terms of installed sites and RFBs, need to be deployed.

Figure 4d reports then the percentage of served users for SFDA, 5G-PCDA, and FFDA vs. the variation of t^{MIN}. As expected, both 5G-PCDA and FFDA are always able to ensure an higher percentage of served users compared with SFDA. For example, 5G-PCDA is able to achieve 100% of served users for $t^{MIN} = \{1, 5, 10\}$ [Mbps], and a percentage higher than 90% for the other values of t^{MIN}. However, we point out that covering a higher percentage of users results in an increase of the monetary costs, as shown in Fig. 4a.

In the next part, we consider the number of installed RRH RFBs vs. the variation of t^{MIN}, as reported in Fig. 5 for all the algorithms.[1] Three considerations hold in this case: (i) the number of T1-RRH RFBs is pretty constant for both SFDA, 5G-PCDA, and FFDA, (ii) the number of T2-RRH RFBs tends to increase with t^{MIN}, (iii) all the algorithms require a similar number of installed T1-RRH RFBs, but different number of installed T2-RRH RFBs. Focusing on (i), the number of deployed T1-RRH RFBs is constant due to the fact that these RFBs are used as "macro cells," in order to cover large portions of territory. Moreover, we recall that there is also a minimum distance of 400 [m] that needs to be ensured between nodes hosting T1-RRH RFBs. Focusing on (ii), T2-RRH RFBs are used to provide capacity to users, i.e., mainly acting as "small cells." Eventually, focusing on (iii) it is clear that, since 5G-PCDA targets the maximization of the number of served users, it requires also an higher number of installed T2-RRH RFBs compared with SFDA. Finally, FFDA also tends to install a large number of T2-RRH RFB, due to the fact that it is un-aware of costs and/or traffic from users.

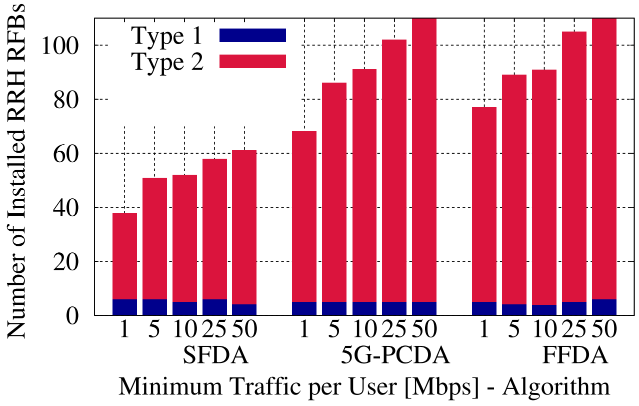


Fig. 5 Number of installed RRH RFBs by SFDA, 5G-PDA, and FFDA vs. the minimum traffic per user t^{MIN} (San Small scenario)

In the following part, we focus on the locations of the installed sites, as well as on the association of users to the installed RRH RFBs. To this aim, Fig. 6 reports the installed sites hosting T1-RRH RFBs or T2-RRH RFBs, the users, and their association to the RRH RFBs. The results from 5G-PCDA with $t^{MIN} = 25$ [Mbps] are reported in the figure. Two considerations hold in this case. First, the number of users served by each T1-RRH RFB is relatively low. Actually, we recall that this type of RFB is used to deploy a "macro cell," whose main goal is to provide coverage of the territory rather than providing extremely high data rates to users. Second, most of users are instead

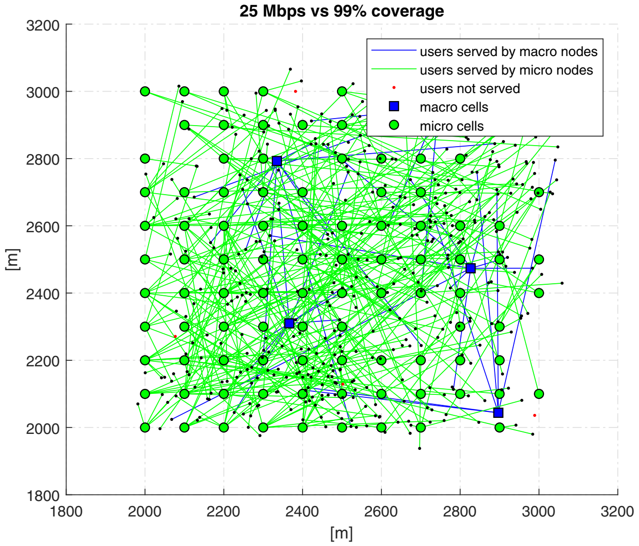


Fig. 6 User to RRH RFB association performed by 5G-PCDA in the SAN Small scenario with t^{MIN}=25 [Mbps]

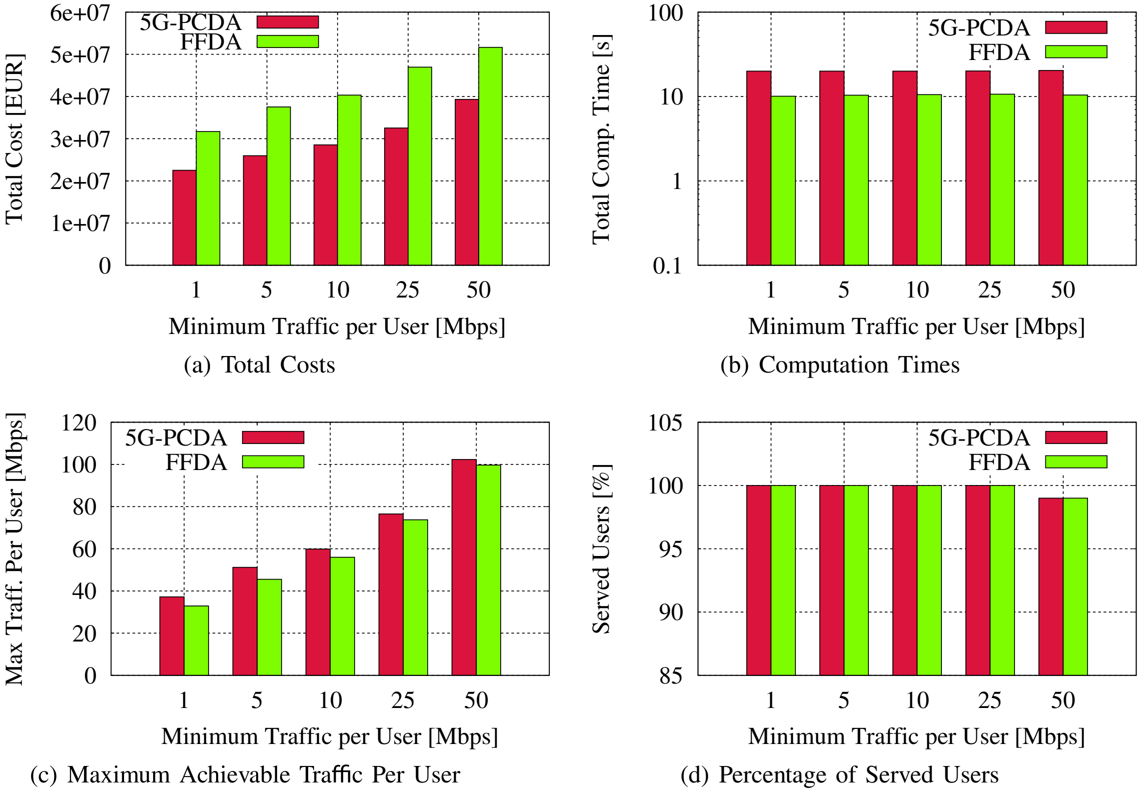
[1] The same analysis was performed on the BBU and MEC RFBs, yielding to the same conclusions (not reported here due to lack of space).

Table 2 Cost Breakdown for SFDA, 5G-PCDA, and FFDA vs. the minimum traffic per user t^{MIN} over the SAN Small scenario

Cost	Algorithm	Min. traffic per user t^{MIN}				
		1 [Mbps]	5 [Mbps]	10 [Mbps]	25 [Mbps]	50 [Mbps]
BBU	SFDA	21922 [€]	27642 [€]	27215 [€]	30722 [€]	30308 [€]
	5G-PCDA	34255 [€]	42175 [€]	44375 [€]	49215 [€]	52735 [€]
	FFDA	38215 [€]	42628 [€]	43508 [€]	50535 [€]	55802 [€]
MEC	SFDA	21922 [€]	27642 [€]	27215 [€]	30722 [€]	30308 [€]
	5G-PCDA	34255 [€]	42175 [€]	44375 [€]	49215 [€]	52735 [€]
	FFDA	38215 [€]	42628 [€]	43508 [€]	50535 [€]	55802 [€]
CHW	SFDA	179018 [€]	240261 [€]	244972 [€]	273238 [€]	287371 [€]
	5G-PCDA	320348 [€]	405146 [€]	428701 [€]	480522 [€]	518210 [€]
	FFDA	362747 [€]	419279 [€]	428701 [€]	494655 [€]	541765 [€]
DHW	SFDA	351123 [€]	471240 [€]	480480 [€]	535920 [€]	563640 [€]
	5G-PCDA	628320 [€]	794640 [€]	840840 [€]	942480 [€]	1016400 [€]
	FFDA	711480 [€]	822360 [€]	840840 [€]	970200 [€]	1062600 [€]
SITE	SFDA	2000000 [€]	2520000 [€]	2480000 [€]	2800000 [€]	2760000 [€]
	5G-PCDA	3120000 [€]	3840000 [€]	4040000 [€]	4480000 [€]	4800000 [€]
	FFDA	3480000 [€]	3880000 [€]	3960000 [€]	4600000 [€]	5080000 [€]

served by the T2-RRH RFBs, which tend to be densely deployed over the territory.

Up to this point, a natural question is then: What is the impact of the single cost components on the total CAPEX? To answer this question, Table 2 reports the cost breakdown for SFDA, 5G-PCDA, and FFDA vs. the variation of t^{MIN}. We recall that the total CAPEX is split in the following components: (i) BBU RFB cost, (ii) MEC RFB cost, (iii) CHW cost, (iv) DHW cost, (v) site cost. Not surprisingly, the site costs heavily impact the total



(a) Total Costs

(b) Computation Times

(c) Maximum Achievable Traffic Per User

(d) Percentage of Served Users

Fig. 7 Performance of the 5G-PCDA and FFDA algorithms vs. the minimum traffic per user t^{MIN} over the SAN Big scenario

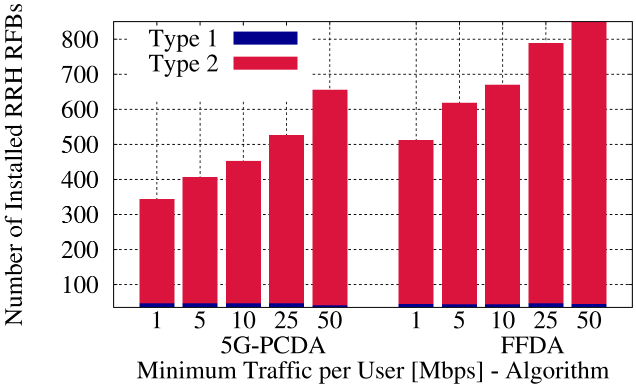


Fig. 8 Number of installed RRH RFBs by 5G-PCDA and FFDA vs. the minimum traffic per user t^{MIN} (San Big scenario)

CAPEX, due to the fact that installing each new site has a large cost for the operator. In addition, the other costs are instead clearly lower. However, all the costs tend to increase with increasing values of t^{MIN}, due to the fact that more resources need to be installed. Finally, the comparison between the the algorithms reveals that SFDA is always less expensive compared with 5G-PCDA and FFDA.

7.2 Results from SAN big scenario

In the last part of our work, we run the SFDA, 5G-PCDA, and FFDA algorithms over the SAN Big scenario. Since this, scenario is much more complex compared with the SAN Small case, the SFDA algorithm, which requires the computation of all the possible configurations in terms of installed T1-RRH RFBs, results to be computationally infeasible (we stopped its execution after several hours, without obtaining any feasible solution). On the other hand, both 5G-PCDA and FFDA are able to retrieve a solution in less than 1 min even in this case. Therefore, we run both 5G-PCDA and FFDA by varying the minimum traffic per user t^{MIN} between 1 and 50 [Mbps]. Figure 7 reports the obtained results, in terms of (i) total costs (Fig. 7a), (ii) computation time (Fig. 7b), (iii) maximum achievable traffic per user (Fig. 7c), and (iv) percentage of served users (Fig. 7d). Interestingly, 5G-PCDA is able to notably reduce the costs compared with FFDA in this case, with a saving in the order of several million euros. This is due to the fact that, contrary to FFDA, 5G-PCDA is able to efficiently to limit the total amount of resources that are installed, while ensuring high performance levels. This is achieved with a slight increase in the computation time compared with FFDA (but still in the order of seconds), which is coupled with a potential higher maximum traffic per user, and a percentage of served users always comparable to FFDA. Overall, the benefits of 5G-PCDA are evident compared with FFDA.

To give more insights, Fig. 8 reports the variation of the number of T1- and T2-RRH RFBs for 5G-PCDA and FFDA. Three considerations hold in this case: (i) the number of T2-RRH RFBs is increasing with t^{MIN} (as expected), (ii) the number of installed T1-RRH RFBs is clearly lower compared with the T2-RRH RFBs, (iii) FFDA installs a consistent higher number of T2-RRH RFBs compared with 5G-PCDA.

Finally, Table 3 reports the breakdown of the costs for the 5G-PCDA algorithm. Interestingly, we can note that the site cost dominates over the other ones, and that the costs are increasing with t^{MIN}. By comparing these results against the ones of the SAN small scenario (see Table 2), we can note an almost 10-fold increase in the different costs components. This is due to the fact that both the dimension of the territory and the number of users in the SAN Big scenario are clearly larger compared with the SAN Small scenario. However, we stress the fact that 5G-PCDA is able to efficiently manage both the increased complexity in the scenario and the provisioning of an adequate service level to users.

8 Conclusions and future work

We have faced the problem of designing a 5G network architecture based on RFBs, with the goal of limiting the total costs while serving the users. We have initially formulated the OPT-5GD problem, which is able to select which 5G nodes and which RFBs have to be installed in the network, in order to serve the users with the amount of

Table 3 5G-PCDA results vs. the minimum traffic per user t^{MIN} over the SAN Big scenario

Metric		Min. traffic per user t^{MIN}				
		1 [Mbps]	5 [Mbps]	10 [Mbps]	25 [Mbps]	50 [Mbps]
Cost	BBU	190362 [€]	218082 [€]	238762 [€]	270882 [€]	324187 [€]
	MEC	190362 [€]	218082 [€]	238762 [€]	270882 [€]	324187 [€]
	CHW	1611162 [€]	1907955 [€]	2129372 [€]	2473275 [€]	3090416 [€]
	DWH	3160080 [€]	3742200 [€]	4176480 [€]	4851000 [€]	6061440 [€]
	Site	17360000 [€]	19880000 [€]	21760000 [€]	24680000 [€]	29520000 [€]

required traffic. After showing that OPT-5GD is NP-Hard, we have proposed the SFDA and 5G-PCDA algorithms to tackle the problem. We have then considered two realistic scenarios located in the city of San Francisco, where the positions of the users and the set of candidate sites to host T1-RRH RFBs are derived from the WeFi app. Our results, show that (i) the total costs are increasing with the minimum amount of served traffic to users t^{MIN}, (ii) SFDA tends to limit the total costs, while 5G-PCDA is able to efficiently compute a solution which tends to serve the largest percentage of users, (iii) the maximum achievable traffic per user is already in the order of dozens of Mbps even for $t^{MIN} = 1$ [Mbps], and (iv) the site costs tend to dominate over the other ones.

As future work, we plan to introduce direct acyclic graphs to model more complex interactions among the RFBs, e.g., one BBU RFB serving multiple RRH RFBs. In addition, we will consider a finer granularity of the RFBs, which can realize simpler functions, and can be run in light execution environments, in line with the current trend of network softwarization. Finally, we plan to investigate the impact of the users mobility, and the uncertainty of user traffic.

Funding information This work has received funding from the Horizon 2020 EU project SUPERFLUIDITY (grant agreement no. 671566).

References

1. View on 5g architecture (version 2.0). (Date last accessed Sep 2017)
2. Galis A, Clayman S, Mamatas L, Loyola JR, Manzalini A, Kuklinski S, Serrat J, Zahariadis T (2013) Softwarization of future networks and services-programmable enabled networks as next generation software defined networks. In: Future networks and services (SDN4FNS), 2013 IEEE SDN for. IEEE, pp 1–7
3. Rost P, Banchs A, Berberana I, Breitbach M, Doll M, Droste H, Mannweiler C, Puente MA, Samdanis K, Sayadi B (2016) Mobile network architecture evolution toward 5g. IEEE Commun Mag 54(5):84–91
4. Rost P, Mannweiler C, Michalopoulos DS, Sartori C, Sciancalepore V, Sastry N, Holland O, Tayade S, Han B, Bega D et al (2017) Network slicing to enable scalability and flexibility in 5g mobile networks. IEEE Commun Mag 55(5):72–79
5. Bianchi G, Biton E, Blefari-Melazzi N, Borges I, Chiaraviglio L, Cruz Ramos P, Eardley P, Fontes F, McGrath MJ, Natarianni L et al (2016) Superfluidity: a flexible functional architecture for 5g networks. Trans Emerg Telecommun Technol 27(9):1178–1186
6. WeFi. http://www.wefi.com
7. Chiaraviglio L, Amorosi L, Cartolano S, Blefari-Melazzi N, Dell'Olmo P, Shojafar M, Salsano S (2017) Optimal superfluid management of 5G networks. In: 3Rd network softwarization, IEEE conference on (IEEE netsoft).IEEE, pp 1–9
8. Chiaraviglio L, D'Andreagiovanni F, Siderotti G, Melazzi NB, Salsano S (2018) Optimal design of 5g superfluid networks: Problem formulation and solutions. In: 21St conference on innovation in clouds, internet and networks (ICIN) 2018
9. Michalopoulos DS, Doll M, Sciancalepore V, Bega D, Schneider P, Rost P (2017) Network slicing via function decomposition and flexible network design. In: Workshop on new radio technologies co-located with IEEE PIMRC. IEEE, Montreal
10. Basta A, Kellerer W, Hoffmann M, Morper HJ, Hoffmann K (2014) Applying nfv and sdn to lte mobile core gateways, the functions placement problem. In: Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, and Challenges, pp 33–38
11. Luizelli MC, Bays LR, Buriol LS, Barcellos MP, Gaspary LP (2015) Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions. In: 2015 IFIP/IEEE international symposium on Integrated network management (IM). IEEE, pp 98–106
12. Yousaf FZ, Loureiro P, Zdarsky F, Taleb T, Liebsch M (2015) Cost analysis of initial deployment strategies for virtualized mobile core network functions. IEEE Commun Mag 53(12):60–66
13. Ananth M, Sharma R (2017) Cost and performance analysis of network function virtualization based cloud systems. In: 2017 IEEE 7th international Advance computing conference (IACC). IEEE, pp 70–74
14. Chen M, Zhang Y, Hu L, Taleb T, Sheng Z (2015) Cloud-based wireless network: virtualized, reconfigurable, smart wireless network to enable 5g technologies. Mob Netw Appl 20(6):704–712
15. Sun S, Kadoch M, Gong L, Rong B (2015) Integrating network function virtualization with sdr and sdn for 4g/5g networks. IEEE Netw 29(3):54–59
16. Rost P, Bernardos CJ, De Domenico A, Di Girolamo M, Lalam M, Maeder A, Sabella D, Wübben D (2014) Cloud technologies for flexible 5G radio access networks. IEEE Commun Mag 52(5):68–76
17. ETSI GS NFV 002 (2014) Network functions virtualisation (NFV); architectural framework v 1.2.1. ETSI
18. Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans Wirel Commun 9(11):3590–3600
19. Hoydis J, Ten Brink S, Debbah M (2013) Massive MIMO in the UL/DL of cellular networks: How many antennas do we need? IEEE J Sel Areas Commun 31(2):160–171
20. Wu J, Zhang Z, Hong Y, Wen Y (2015) Cloud radio access network (c-ran): a primer. IEEE Netw 29(1):35–41
21. D'Andreagiovanni F, Caire G (2016) An unconventional clustering problem: user service profile optimization. In: 2016 IEEE international symposium on Information theory (ISIT)IEEE, pp 855–859
22. Kellerer H, Pferschy U, Pisinger D (2004) Knapsack problems. Springer, Berlin
23. Martello S, Toth P (1990) Knapsack problems: algorithms and computer implementations. Wiley, New York
24. Malandrino F, Chiasserini C-F, Kirkpatrick S (2018) Cellular network traces towards 5g: usage, analysis and generation. IEEE Trans Mob Comput 17(3):529–542
25. Chiaraviglio L, Blefari-Melazzi N, Chiasserini CF, Iatco B, Malandrino F, Salsano S (2017) An economic analysis of 5G Superfluid networks. In: 2017 IEEE 18th international conference on High performance switching and routing (HPSR). IEEE, pp 1–7
26. Coffman EG Jr, Csirik J, Galambos G, Martello S, Vigo D (2013) Bin packing approximation algorithms: survey and classification. Handbook of combinatorial optimization, pp 455–531