Optimal Design of 5G Superfluid Networks: Problem Formulation and Solutions

Luca Chiaraviglio[1,2], Fabio D'Andreagiovanni[3,4], Giulio Sidoretti[2], Nicola Blefari-Melazzi[1,2], Stefano Salsano[1,2],
1) Consorzio Nazionale Interunivesitario per le Telecomunicazioni, Italy,
2) EE Department, University of Rome Tor Vergata, Italy, email: luca.chiaraviglio@uniroma2.it
3) National Center for Scientific Research (CNRS), France,
4) Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc
UMR 7253, CS 60319, 60203 Compiègne, France, email: d.andreagiovanni@hds.utc.fr

***Abstract*—The forthcoming 5G technology foresees the exploitation of solutions able to increase both the flexibility and the scalability of the network. In line with the current trend of softwarization, in this work we face the problem of designing a 5G network from the outcome of the Horizon 2020 project Superfluidity. The core of the project is the definition of a 5G converged architecture based on virtual entities, called Reusable Functional Blocks (RFBs), which can be run on different HardWare (HW) and SoftWare (SW) execution environments. The exploitation of RFBs allows to achieve the required level of flexibility required by 5G. After optimally formulating the problem of minimizing the total installation costs of a SuperFluid network composed of RFBs and physical 5G nodes, we propose a new algorithm, called SFDA, to practically tackle the problem. Our results, obtained over a representative case study, show that SFDA is able to solve the problem in a reasonable amount of time, returning solutions very close to the optimum. In addition, we clearly show the trade offs that emerge between the need of providing a service level to users (in terms of downlink traffic or coverage) and the total costs incurred to install the elements of the network.**

I. Introduction

According to the 5G Public Private Partnership (PPP), the forthcoming 5G tecnology 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 will be 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 to different 5G nodes. The core of the project is the definition of a cloud-based 5G converged solution, in which softwarized 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: How to minimize the installation costs of a 5G SF architecture, while still guaranteeing the 5G service to users? How to optimally formulate the problem? Is it possible to design a smart algorithm to design the 5G SF network? 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 a 5G SF 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 an efficient heuristic, called SuperFluid Design Algorithm (SFDA), to reduce the computation times while ensuring a good performance to users;
- we solve the mathematical formulation on a realistic scenario, comparing the performance of the SFDA algorithm with that of a state-of-the-art optimization solver and analyzing the corresponding trade-offs.

To the best of our knowledge, none of the previous works has conducted a similar analysis. The closest paper to our work is [6], in which the authors have targeted the efficient management of the RFBs in a SF network, with the goal of maximizing the traffic per user or the number of used nodes. However, the work in [6] 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 [6] 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 [6]. In particular, the elements installed during the design phase, which are selected by this work, can be used as input for the management one.

Our results clearly show that the costs for designing the 5G SF network can be minimized, while guaranteeing an adequate Quality of Service (QoS) perceived by users. In addition, the proposed SFDA algorithm is able to identify solutions that are very close to the optimum, while being able to limit the computation times to some seconds in the worst case. 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 investigating the impact of the size of the scenario on the results. We leave the evaluation of these aspects as future work.

The rest of the paper is organized as follows. The SF architecture is overviewed in Sec. II. The optimal formulation is detailed in Sec. III. Sec. IV includes the description of the SFDA algorithm. Sec. V details the scenario and the parameter settings. The performance of the optimal formulation and of the SDFA algorithm is reported in Sec. VI. Sec. VII overviews the related works. Finally, conclusions are drawn in Sec. VIII.

II. SUPERFLUID 5G ARCHITECTURE

We report here a brief overview of the 5G SF 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 advantage 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 [7], 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 SF 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

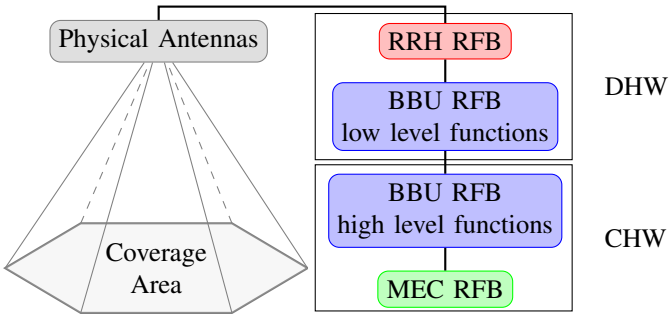


Fig. 1. Scheme of a 5G SuperFluid node serving an area.

Massive User Multiple Input Multiple Output (MU-MIMO) technology [8], [9]. 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. Fig. 1 reports a scheme of 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 (C-RAN) paradigm [10]. 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 a 5G SF network

composed of RFBs.

III. 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 RFBs 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 RFBs and MEC RFBs placement. In particular, we introduce the set B and the set M to store the BBU RFBs types and the MEC RFBs 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:

$$\sum_{n_1 \in N} \sum_{n_2 \in N} v_{n_1 n_2 m}^{MEC} \leq a_m^{MEC} \qquad m \in M \tag{8}$$

where $v_{n_1 n_2 m}^{MEC}$ 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_m^{MEC} 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:

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

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:

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

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:

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

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., [11]) by introducing one *continuous* variable $\phi_{un} = t_u \, x_{un}$ and the four linear inequalities:

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

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

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

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

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

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

The linear version of constraint (11) is then:

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

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_r^{RRH}. We express this condition with the following constraint:

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

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

$$\theta_{unr} \geq 0 \tag{15a}$$
$$\theta_{unr} \leq x_{un} \tag{15b}$$
$$\theta_{unr} \leq y_{nr}^{RRH} \tag{15c}$$
$$\theta_{unr} \geq x_{un} + y_{nr}^{RRH} - 1 \tag{15d}$$

The linear version of constraint (14) is then:

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

We then introduce the input parameter CAP_m^{MEC}, 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:

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

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 (11), we first use the linearization variables introduced in (13), imposing $\phi_{un_1} = t_u x_{un_1}$; then we face the resulting product of variables $\phi_{un_1} v_{n_1 n_2 m}^{MEC}$, which can be linearized by introducing a new continuous variable $\varphi_{un_1 n_2 m} = \phi_{un_1} v_{n_1 n_2 m}^{MEC}$ and the following four constraints:

$$\varphi_{un_1 n_2 m} \geq 0 \tag{18a}$$
$$\varphi_{un_1 n_2 m} \leq \text{CAP}_u^{\max}\, v_{n_1 n_2 m}^{MEC} \tag{18b}$$
$$\varphi_{un_1 n_2 m} \leq \phi_{un_1} \tag{18c}$$
$$\varphi_{un_1 n_2 m} \geq \phi_{un_1} - (1 - v_{n_1 n_2 m}^{MEC})\, \text{CAP}_u^{\max} \tag{18d}$$

The linear version of constraint (17) is then:

$$\sum_{u \in U} \sum_{n_1 \in N} \varphi_{un_1 n_2 m} \leq \text{CAP}_m^{MEC} \sum_{n_1 \in N} v_{n_1 n_2 m}^{MEC}, \qquad n_2 \in N, m \in M \tag{19}$$

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:

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

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

$$v_{n_1 n_2 m}^{MEC} \leq y_{n_1 r}^{RRH} \qquad r \in R, n_1, n_2 \in N, m \in M \tag{21}$$
$$v_{n_1 n_2 b}^{BBU} \leq y_{n_1 r}^{RRH} \qquad r \in R, n_1, n_2 \in N, b \in M \tag{22}$$

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_r^{SITE} 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_b^{BBU} and c_m^{MEC} 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:

$$\min \sum_{n \in N} \sum_{r \in R} \left(c_r^{SITE} + c^{CH} + c^{DH}\right) y_{rn}^{RRH} + \\ + \sum_{n_1 \in N} \sum_{n_2 \in N} \left(\sum_{b \in B} c_b^{BBU} v_{n_1 n_2 b}^{BBU} + \sum_{m \in M} c_m^{MEC} v_{n_1 n_2 m}^{MEC} \right) \tag{24}$$

Users to RRH RFBs assignment:	Eq. (1), (2)
RRH RFBs 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. (13), (16)
Maximum MEC RFB capacity	Eq. (19)
RRH RFBs conflict constraint	Eq. (20)
MEC/BBU RFBs placement constraints	Eq. (21), (22)
Minimum Traffic Constraints	Eq. (23)
Linearization Constraints	
Eq. $(12a - 12d), (15a - 15d), (18a - 18d)$	

(25)

Under variables: $x_{un} \in \{0,1\}$, $t_u \geq 0$, $y_{nr}^{RRH} \in \{0,1\}$, $v_{n_1 n_2 b}^{BBU} \in \{0,1\}$, $v_{n_1 n_2 m}^{MEC} \in \{0,1\}$, $\phi_{un_1} \geq 0$, $\theta_{unr} \geq 0$, $\varphi_{un_1 n_2 m} \geq 0$.

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

Algorithm 1 Pseudocode of the SuperFluid Design Algorithm (SFDA)

1: **Input:** N, U, a_r^{RRH}, a_b^{BBU}, a_m^{MEC}, CAP_{run}, t^{MIN}, δ, order_type
2: **Output:** y_{nr}^{RRH}, $v_{n_1n_2b}^{BBU}$, $v_{n_1n_2b}^{MEC}$, x_{un}
3: tot_cost_best_conf=Inf;
4: all_conf=comp_conf(N,a_r^{RRH},$r=1$);
5: **for** curr_conf in all_conf **do**
6: tot_RRH_RFB=0;
7: u_cand_served= comp_cand_served_u(curr_conf, order_type, U, t^{MIN});
8: n_sorted=sort_RRH_RFB(u_cand_served, curr_conf, $r=1$);
9: curr_u_to_serve=U;
10: **for** n in n_sorted **do**
11: u_assoc=associate_u(n, curr_u_to_serve, t^{MIN}, $r=1$);
12: curr_u_to_serve=remove_served_u(U, u_assoc);
13: **end for**
14: n_sorted=sort_RRH_RFB(curr_u_to_serve, N, $r=2$)
15: **for** n in n_sorted **do**
16: **if** check_tot_u_served(u_assoc,δ)==false) **then**
17: **if** (check_conf(curr_conf,n)==true)&& (tot_RRH_RFB< $a_{r=2}^{RRH}$) **then**
18: tot_RRH_RFB=tot_RRH_RFB+1;
19: curr_conf=add_RRH_RFB(curr_conf, n, $r=2$);
20: u_assoc=associate_u(n, curr_u_to_serve, t^{MIN}, $r=2$);
21: curr_u_to_serve=remove_served_u(U, u_assoc);
22: **end if**
23: **end if**
24: **end for**
25: MEC_BBU_RFB_conf=assign_BBU_MEC_RFB(curr_conf, u_assoc, a_b^{BBU}, a_m^{MEC}, t^{MIN});
26: tot_cost=comp_tot_cost(curr_conf);
27: **if** (tot_cost<tot_cost_best_conf)&& (check_tot_u_served(u_assoc,δ)==true) **then**
28: tot_cost_best_conf=tot_cost;
29: [y_{nr}^{RRH}, $v_{n_1n_2b}^{BBU}$, $v_{n_1n_2b}^{MEC}$, x_{un}]= save_conf(curr_conf, u_assoc, t^{MIN});
30: **end if**
31: **end for**

IV. Algorithm Description

We detail here the main steps of the SuperFluid Design Algorithm (SFDA). We design the 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 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 to 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, 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}, 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 precence 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), 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

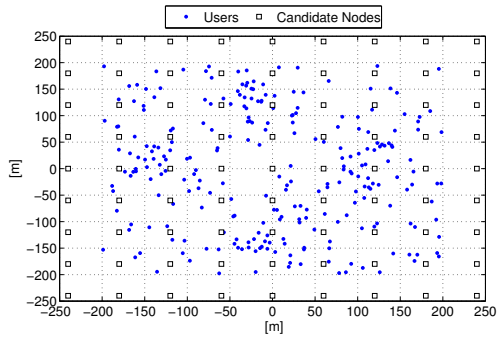


Fig. 2. Candidates nodes and users positions in the considered scenario.

TABLE I
INPUT PARAMETERS

Parameter	Value
$\|U\|$	258
$\|N\|$	81
U_r^{MAX}	126 (T1-RRH RFB) 42 (T2-RRH RFB)
a_r^{RRH}	5 (T1-RRH RFB) 81 (T2-RRH RFB)
a_b^{BBU}	5 (T1-BBU RFB) 81 (T2-BBU RFB)
a_m^{MEC}	5 (T1-MEC RFB) 81 (T2-MEC RFB)
CAP_{run}	Model from Marzetta [8] with input parameters from [6].
CAP_r^{RRH}	30 [Gbps] (T1-RRH RFB) 10 [Gbps] (T2-RRH RFB)
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}	120 [k€] (T1-RRH RFB) 40 [k€] (T2-RRH RFB)
c^{CH}	4711 [€]
c^{DW}	9240 [€]
c_b^{BBU}	1307 [€] (T1-BBU RFB) 440 [€] (T2-BBU RFB)
c_m^{MEC}	1307 [€] (T1-BBU RFB) 440 [€] (T2-BBU 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 the assignment of each user to each RRH RFB.

V. SCENARIO AND PARAMETERS SETTINGS

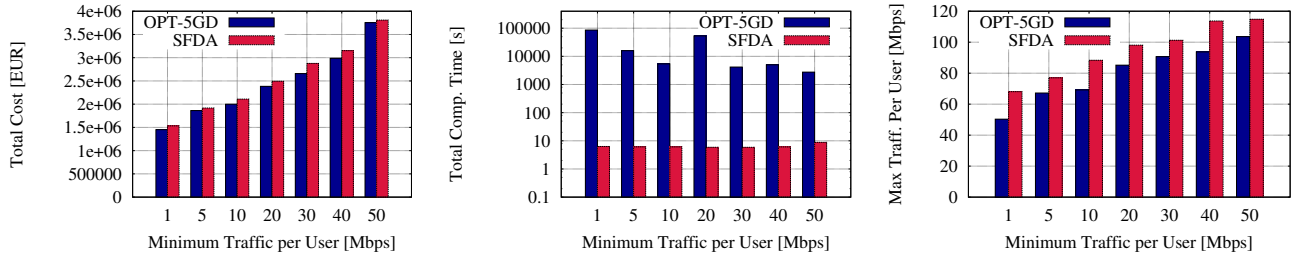
We consider an area of size 500×500 [m^2], where a set of around 260 users are deployed. More in depth, 70% of users are placed in the surroundings of four hot spots, while 30% are randomly placed over the area. In addition, we consider as candidate sites to install the 5G nodes the points at the interesections of a square grid, with a distance of 60 [m] between any two consecutive points. Fig. 2 reports the positions of the candidates sites ans well the users over the considered area.

Given this scenario, we set the input parametes, which are summarized in Tab. I. Unless otherwise specified, we adopt a similar setting of input parameters as in [6]. More in detail the T1-RRH RFB is able to serve more users compared to the T2-RRH RFB. In addition, we consider a low number of available T1 RFBs, and a large number of T2 RFBs. In particular, the number of T2 RFBs is set equal to the cardinality of the set of candidate sites. Focusing then on the downlink capacity model, we adopt the same model of Marzetta [8]. We refer the reader to [6] for a detailed description of the parameters adopted for this model. Moreoveor, the compatibility matrix of possible configurations $CONF_r$ is set in accordance to the following rules: i) each pair of T1-RRH RFBs 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 in each site. Focusing then on the cost, we assume that the site installation costs are higher for the nodes hosting T1-RRH RFBs compared to 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) [12], which has to be properly dimensioned. Note that, in [12] 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.

VI. PERFORMANCE EVALUATION

We evaluate the OPT-5GD formulation and the SFDA algorithm over the considered scenario. We code the problem formulation by adopting the optimization libraries provided by CPLEX v. 12.7.1. We then run the formulation on a high performance computing cluster provided by the Azure Cloud, with 8 cores (each of them dual threaded) and 56 [GB] of RAM. In addition, we code the SFDA algorithm in Matlab, and we run it on a notebook equipped with 2 cores Intel Core i7 at 2.8 [GHz] and 8 [GB] of RAM.

We initially evaluate the impact of varying the minimum amount of traffic t^{MIN} between 1 [Mbps] and 50 [Mbps]. Moreover, we initially set the percentage of 5G users δ equal to 95%. Fig. 3(a) 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 three times when t^{MIN} passes from 1 [Mbps] to 50 [Mbps]. The relatively small increase of the total costs compared to 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



(a) Total Costs (b) Computation Times (c) Maximum Achievable Traffic Per User

Fig. 3. Performance of the OPT-5GD formulation and SFDA algorithm vs. the minimum traffic per user t^{MIN}.

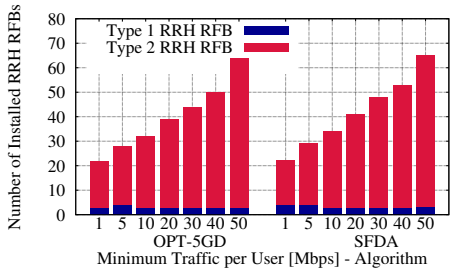
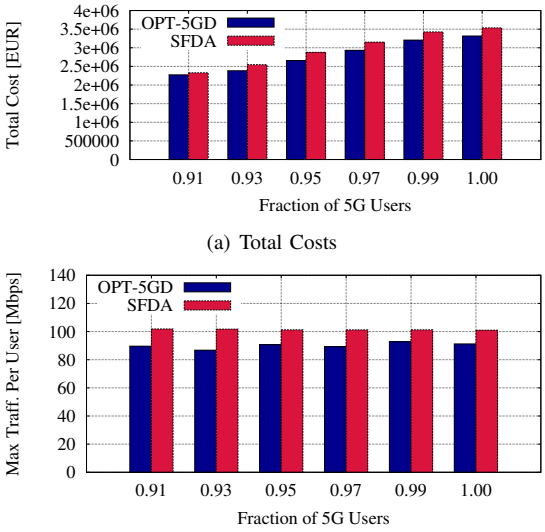


Fig. 4. Number of installed RRH RFBs by the OPT-5GD formulation and the SFDA algorithm vs. the minimum traffic per user t^{MIN}.



(a) Total Costs

(b) Maximum Achievable Traffic Per User

Fig. 5. Performance of the OPT-5GD formulation and the SFDA algorithm vs. the percentange of covered 5G users δ.

that SFDA requires a relatively small amount of additional costs compared to OPT-5GD.

In the following, we consider the impact of computation times, as reported in Fig. 3(b). As expected, the OPT-5GD requires a not negligible time to be solved. Moreover, the computation time of OPT-5GD tends to increase when t^{MIN} is decreased, due to the fact that in this case there are more possibilities in placing the 5G nodes and selecting the RFBs in order to serve the users. On the other hand, SFDA is always able to retrieve a solution in few seconds. In addition, the optimal formulation may become very challenging to be solved in a larger scenario, thus motivating us for the adoption of the SFDA solution in this case.

We then consider the maximum amount of traffic that can be served for each user. In particular, given the output of OPT-5GD 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. In a similar way, we have computed the maximum traffic also from the ouput of SFDA. Fig. 3(c) reports the obtained results. Interestingly, both the solutions are able to provide a large throughput to users, i.e. more than 40 [Mbps], even when $t^{MIN} \leq 10$ [Mbps]. This is due to the fact that the capacity of the installed RFBs is able to ensure large requests of traffic of 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 the next part, we consider the number of installed RRH RFBs vs. the variation of t^{MIN}, as reported in Fig. 4.[1] Three considerations hold in this case: i) the number of T1-RRH RFBs is pretty constant, and ii) the number of T2-RRH RFBs tends to increase with t^{MIN}, iii) both OPT-5GD and SFDA require a similar number of installed 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".

In the last part of our work, we set $t^{MIN} = 30$ [Mbps] and we vary the fraction of 5G users δ between 0.91 and 1.00. We then run again OPT-5GD and SFDA over these scenarios. Fig. 5 reports the obtained results. As expected, the total costs (Fig. 5(a)) are increasing with δ. Again, we can see that SFDA performs very close to OPT-5GD. Interestingly, in this case the maximum amount of traffic (Fig. 5(b)) is pretty independent from δ. This is due to the fact that this metric is more related to t^{MIN} (i.e., as reported in Fig. 3(c)) than to δ.

[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).

VII. 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 [13], the author focus on the concept of network function decomposition in conjunction with its relation to network slicing. Both [3] and [13] 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 [14] 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 [15], the considered VNFs are firewalls, load balancers, 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 [16], the authors focus on a 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 [17] study the influence of NFV on CAPEX of cloud based networks and compare it with traditional implementation without NFV in few example scenarios. However, no general optimization models are considered.

VIII. Conclusions and Future Work

We have faced the problem of minimizing the total costs in a 5G SF network architecture. We have initially formulated the OPT-5GD problem, which is able to select which nodes and which RFBs have to be installed in the network, in order to serve the users with the amount of required traffic. In the following, we have proposed the SFDA algorithm to practically tackle the problem. Our results, obtained over a representative scenario, show that: i) the total costs are increasing with the minimum amount of served traffic to users t^{MIN}, ranging from around 1.5×10^6 [€] when $t^{MIN} = 1$ [Mbps] to more than 3.5×10^6 [€] when $t^{MIN} = 50$ [Mbps], ii) SFDA performs very close to the optimal solution, while reducing the computation times to less than 10 [s], iii) the maximum achievable traffic per user is already larger than 50 [Mbps] for $t^{MIN} \geq 5$ [Mbps] and iv) when the fraction of 5G users is increased, an increase in the total costs is incurred.

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. We plan also to compare SFDA with other VNFs placement algorithms. Eventually, we will target the optimization of the installation costs and the maximization of the number of served users. Finally, 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.

Acknowledgments

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] A. Galis, S. Clayman, L. Mamatas, J. R. Loyola, A. Manzalini, S. Kuklinski, J. Serrat, and T. Zahariadis, "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*, pp. 1–7, IEEE, 2013.

[3] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5g," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, 2016.

[4] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, *et al.*, "Network slicing to enable scalability and flexibility in 5g mobile networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, 2017.

[5] G. Bianchi, E. Biton, N. Blefari-Melazzi, I. Borges, L. Chiaraviglio, P. Cruz Ramos, P. Eardley, F. Fontes, M. J. McGrath, L. Natarianni, *et al.*, "Superfluidity: a flexible functional architecture for 5g networks," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 9, pp. 1178–1186, 2016.

[6] L. Chiaraviglio, L. Amorosi, S. Cartolano, N. Blefari-Melazzi, P. Dell'Olmo, M. Shojafar, and S. Salsano, "Optimal Superfluid Management of 5G Networks," in *3rd Network Softwarization, IEEE Conference on (IEEE NetSoft)*, pp. 1–9, IEEE, 2017.

[7] "Etsi gs nfv 002: Network functions virtualisation (nfv); architectural framework, v 1.2. 1," *ETSI, December*, 2014.

[8] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[9] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?," *IEEE Journal on selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, 2013.

[10] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (c-ran): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, 2015.

[11] F. D'Andreagiovanni and G. Caire, "An unconventional clustering problem: user service profile optimization," in *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 855–859, IEEE, 2016.

[12] L. Chiaraviglio, N. Blefari-Melazzi, C. F. Chiasserini, B. Iatco, F. Malandrino, and S. Salsano, "An economic analysis of 5G Superfluid networks," in *High Performance Switching and Routing (HPSR), 2017 IEEE 18th International Conference on*, pp. 1–7, IEEE, 2017.

[13] D. S. Michalopoulos, M. Doll, V. Sciancalepore, D. Bega, P. Schneider, and P. Rost, "Network Slicing via Function Decomposition and Flexible Network Design," in *Workshop on New Radio Technologies co-located with IEEE PIMRC, Montreal, Canada*, IEEE, October 2017.

[14] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying nfv and sdn to lte mobile core gateways, the functions placement problem," in *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, and Challenges*, pp. 33–38, 2014.

[15] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," in *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pp. 98–106, IEEE, 2015.

[16] F. Z. Yousaf, P. Loureiro, F. Zdarsky, T. Taleb, and M. Liebsch, "Cost analysis of initial deployment strategies for virtualized mobile core network functions," *IEEE Communications Magazine*, vol. 53, no. 12, pp. 60–66, 2015.

[17] M. Ananth and R. Sharma, "Cost and Performance Analysis of Network Function Virtualization Based Cloud Systems," in *Advance Computing Conference (IACC), 2017 IEEE 7th International*, pp. 70–74, IEEE, 2017.