Optimal Superfluid Management of 5G Networks

Luca Chiaraviglio,[1,2] Lavinia Amorosi,[3] Stefania Cartolano,[3] Nicola Blefari-Melazzi,[1,2]
Paolo Dell'Olmo,[3] Mohammad Shojafar,[1] Stefano Salsano[1,2]
1) CNIT, Italy, email {name.surname}@cnit.it
2) EE Department, University of Rome Tor Vergata, email {name.surname}@uniroma2.it
3) DSS Department, University of Rome Sapienza, Rome, Italy, email {name.surname}@uniroma1.it

Abstract—**We consider the problem of evaluating the performance of a 5G network based on reusable components, called Reusable Functional Blocks (RFBs), proposed by the Horizon 2020 SUPERFLUIDITY project. RFBs allow a high level of flexibility, agility, portability and high performance. After formally modelling the RFB entities and the network physical nodes, we optimally formulate the problem of maximizing different Key Performance Indicators (KPIs) on an RFB-based network architecture, in which the RFBs are shared among the nodes, and deployed only where and when they are really needed. Our results, obtained by solving the proposed optimization problem over a simple yet representative scenario, show that the network can be managed in a very efficient way. More in depth, the RFBs are placed into the nodes in accordance with the amount of requested traffic from users and the specific pursued KPI, e.g., maximization of user throughput or minimization of the number of used nodes. Moreover, we evaluate the relationship between the capacity of each node and the number of RFBs deployed on it.**

I. Introduction

The Internet is becoming an ever increasing pervasive technology. According to different studies, new services like High Definition (HD) videos, tactile applications [1], Internet of Things (IoT) [2] and extremely low delay applications will dominate the scene in the forthcoming years. In addition to this, the number of users will continue to notably increase, especially for countries located in the Far East. As a result, the network itself will have to evolve from a monolithic architecture towards a converged, flexible and high performance solution. To this end, new paradigms like Network Function Virtualization (NFV) [3], Mobile Edge Computing (MEC) [4], Cloud Radio Access Networks (C-RAN) [5], and Massive Multiple Input Multiple Output (Massive MIMO) [6] have been proposed in the last years. Moreover, several initiatives are currently devoted to the design of 5G networks [7], which are expected to turn into reality by 2020.

In this context, the SUPERFLUIDITY project [8] (funded by the EU through the H2020 program) aims to design a 5G network architecture having key features, such as: i) high flexibility, ii) agility, iii) portability, and, iv) high performance. The core of the project is the definition of a *superfluid* approach, meaning that network functions and services are decomposed into *reusable* components, denoted as Reusable Functional Blocks (RFBs), which are deployed on top of physical nodes. More in depth, RFBs have notable features, including: i) RFBs chaining, in order to implement more complex functionality and provide the required service to user; ii) platform independence, i.e., RFBs can be realized via software functions, and can run on several hardware solutions; and, iii) high flexibility and performance, thanks to the fact that RFBs can be deployed where and when they are really needed (hence the *superfluid* attribute of the architecture). The RFB concept is a generalization of the Virtual Network Function (VNF) concept proposed by ETSI [9]. In particular, RFBs can be arbitrarily decomposed in other RFBs, while VNFs in the ETSI model cannot be composed in other VNFs. Moreover, the RFBs can be mapped into different SoftWare (SW) and HardWare (HW) execution environments (see [8]), while the ETSI model focuses on mapping VNFs into Virtual Machines (or Containers) in traditional cloud infrastructures.

In this context, several questions arise, like: Is it possible to efficiently manage a 5G superfluid network based on RFBs? How to model the network and services to evaluate the user performance? How to optimally map the RFBs on the network nodes under different Key Performance Indicators (KPIs)? The answer to these questions is the goal of the paper. Specifically, we consider a NFV-based 5G architecture to model the needed components in terms of RFBs and the infrastructure resources in terms of physical nodes and HW features. We then optimally formulate the problem of managing a set of RFBs in order to serve the users of a 5G network with a high definition video distribution service. Our results, obtained over a simple yet representative case study, allow to evaluate the performance of the NFV based architecture for 5G networks. Moreover, we point out that the proposed approach is a first step towards a more comprehensive solution. Specifically, in this work we focus on RFBs types that can be mapped in VNFs of the ETSI model. The full evaluation of other RFBs features (such as the decomposition of RFBs in smaller RFBs, and the mapping of RFBs to different software environments) will be two interesting branches of future research.

The rest of the paper is organized as follow. The description of the 5G architecture under investigation is reported in Sec. II. The models of the 5G network resources and RFBs are reported in Sec. III. The optimal formulations under different KPIs are detailed in Sec. IV. Sec. V then describes the 5G scenario under investigation. The performance evaluation of the optimal formulations is reported in Sec. VI. Finally, Sec. VII concludes our work.

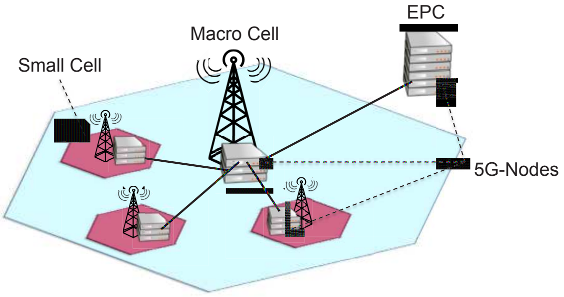


Fig. 1. Physical system infrastructure.

II. Architecture Description

The 5G network model considered in this work is composed of a set of nodes, a set of links and a set of users. The nodes are used to deploy either small cells, macro cells, or to realize the core network elements of the so called Evolved Packet Core (EPC). Each node is connected to the rest of the network by means of a path of physical links. Each user can be connected to the network by means of a cell (either a macro cell or a small cell). For simplicity, the EPC elements are collapsed in a single site in our model.

Fig. 1 reports an example of the considered physical system infrastructure, which is composed of different small cells sites, one macro cell site and one EPC site. In this scenario, each site corresponds to a 5G node. The figure reports also the coverage areas of the cells (which are represented by hexagonal layouts for the sake of simplicity). The service area, i.e., the area where the users are located, is assumed to be overlapped with the coverage area of the macro cell.

Each 5G node is able to host different RFBs. An RFB performs specific tasks in the network architecture, such as processing the video to users, or performing networking and physical layer tasks. In addition, each RFB consumes an amount of physical resources on the hosting 5G node. As physical resources we consider the *processing capacity* (that will be simply denoted as *capacity* further on) and the *memory occupation* (in short denoted as *memory*).

The following RFBs types are taken into consideration in this work:

- Mobile Edge Computing (MEC) RFB;
- Base Band Unit (BBU) RFB;
- Resource Radio Head (RRH) RFB.

We then briefly describe each RFB type in more detail.

MEC RFB. This module is responsible for providing the HD video distribution service to users. A practical example of a MEC RFB is a cache serving a set of videos to users. In general, this module is able to serve an amount of traffic, and consequently a subset of the users spread over the service area. Clearly, the maximum amount of traffic that can be served depends on the amount of resources that are made available to the RFB by the physical node hosting it.

BBU RFB. This module acts as an interface between the MEC RFB and the RRH RFB. Specifically, the BBU RFB exchanges an amount of IP traffic with the MEC module, and a baseband signal with the RRH one. Similarly to the MEC case, also this module is characterized by an amount of consumed resources to provide the RFB functionality.

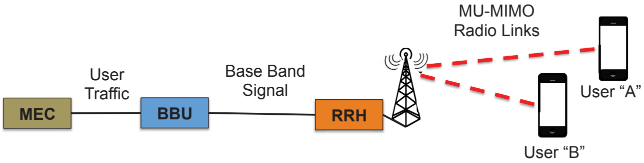


Fig. 2. RFBs relationship and exchanged information.

RRH RFB. This module performs physical layer operations. Specifically, the RRH module handles a set of Radio Frequency (RF) channels with users and the corresponding baseband channels with the BBU RFB. The amount of resources required by this module depends on the type of deployed cell (either a small cell or a macro cell).

In the following, we focus on the interactions among the RFBs. In our context, the RFBs are organized in logical chains. Specifically, each MEC RFB is logically connected to a BBU RFB, which, in turn, is connected to a RRH RFB and consequently to a set of users. Fig. 2 reports an example of RFBs chain and the exchanged information between the modules and the users. In addition, the connection between a pair of RFBs in the chain can be direct, i.e., both RFBs are located on the same physical 5G node, or indirect, i.e., the RFBs are located on two separate nodes. In this latter case, the information flows on an external physical link. Finally, RRH RFBs are able to setup a radio link with users, by exploiting the Multi User Multiple Input Multiple Output (MU-MIMO) technology.

Focusing then on the placement of RFBs in the 5G nodes, the RRH RFBs can be placed only in nodes connected to the antennas of the Radio Access Network (RAN). On the contrary, BBU RFBs can be pooled in other nodes (i.e., by exploiting the Cloud-RAN paradigm). Finally, MEC RFBs can be potentially deployed in every node of the network.

The key feature of the considered NFV-based 5G system is that the RFBs are fully virtualized resources. Specifically, the RFBs can be dynamically moved across the nodes to satisfy the KPIs of the network operator, e.g., the maximization of the user performance or the minimization of the number of used 5G nodes.

III. 5G Node Model and RFBs Models

We then move our attention to a more formal modeling of the 5G nodes and of the RFB types. Let us denote with $\mathcal{N}$ the set of 5G nodes and with $\mathcal{U}$ the set of users, respectively. In the following, we focus on a generic node $i \in \mathcal{N}$ and an RFB chain entirely deployed on it.

A. 5G Node Model

We assume that each node is composed of a Dedicated HardWare (DHW) and a Commodity HardWare (CHW) part. More in depth, the DHW part hosts RFB functionalities requiring intensive and HW specific operations. Such operations include the RRH functions and the BBU functions involving RF and baseband processing tasks. On the other hand, the CHW part of the node is used to host RFB functionalities requiring basic processing tasks (e.g., processing of IP packets or of video traffic), which are performed by the MEC RFBs and the processing functions of BBU RFBs. Fig. 3 reports a scheme of a 5G node, including the CHW and the DHW parts. The node in the example hosts one MEC RFB in the CHW, one RRH RFB in the DHW and one BBU RFB split between the CHW and DHW parts.

Each RFB then consumes an amount of physical resources on the hosting 5G node. Focusing on DHW, we assume that the RFBs require purely capacity resources. More formally, let us denote with δ_i^{RRH} the amount of capacity required by an RRH RFB hosted in node i. In addition, let us denote with δ_i^{BBU} the amount of capacity required by the baseband tasks of BBU RFB hosted at node i. Clearly, the total amount of resources required by the RFBs has to be lower than the DHW installed capacity B_i^{DHW}:

$$\delta_i^{RRH} + \delta_i^{BBU} \leq B_i^{DHW} \tag{1}$$

Focusing then on the CHW part of the node, we assume that the resources required by RFBs are constrained by both the capacity (i.e. maximum utilization of the CPUs) and the memory occupation. More formally, let us denote with C_i^{MEC} and C_i^{BBU} the amount of processing capacity required by the MEC RFB and the BBU RFB on node i, respectively. Similarly, we denote with M_i^{MEC} and M_i^{BBU} the amount of memory required by the MEC RFB and the BBU RFB, respectively. These resources are then bounded by the maximum CPU utilization (C_i^{CHW}) and the maximum memory utilization (M_i^{CHW}) of the node:

$$C_i^{MEC} + C_i^{BBU} \leq C_i^{CHW} \tag{2}$$

$$M_i^{MEC} + M_i^{BBU} \leq M_i^{CHW} \tag{3}$$

The following subsections then detail the modeling of each RFB type and of the associated resources consumed on the node.

B. RRH RFB Model

The RRH RFB module is responsible for serving a set of users with radio resources. Specifically, the following features are adopted: Multi User Multiple Input Multiple Output (MU-MIMO), frequency reuse, Time Division Duplex (TDD), and Orthogonal Frequency Division Multiplexing (OFDM). Specifically, the RRH RFB placed on the node is connected to an array of physical antennas. Moreover, we assume that each user device is equipped with a single antenna. Similarly to [10], we assume that the number of installed antennas is larger than the number of users served by the cell (either a small cell RRH or a macro cell RRH). In this way, we can rely on [10] to easily compute both the maximum number of served users per RRH RFB as well as the radio link capacity provided to each user.[1]

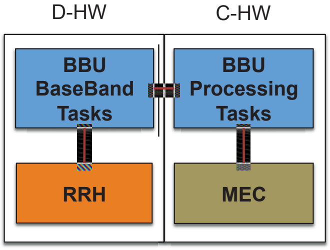


Fig. 3. 5G-Node architecture. The Commodity Hardware (CHW) hosts MEC RFBs and BBU processing tasks. The Dedicated Hardware (DHW) hosts BBU baseband tasks and RRH RFBs.

Let us denote with U^{max} the maximum number of users that can be served by a single RRH RFB located at node i (to ease the notation we do not distinguish between macro cell and small cell sites for the moment). U^{max} is bounded by the reverse link constraint of [10]:

$$U^{max} = \left(\frac{\tau T_u}{T_d}\right) \tag{4}$$

where τ is the number of OFDM symbols used for pilots, T_u is the useful symbol duration (which can be expressed as $T_u = 1/\delta_f$, where δ_f is the subcarrier spacing), and T_d is the largest possible delay spread. Let us denote with T_g and T_s the guard interval and the symbol interval, respectively. More in depth, the symbol interval is expressed as $T_s = T_c/N_{OFDM}$, where T_c is the coherence time and N_{OFDM} is the number of OFDM symbols. In addition, T_g is expressed as $T_g = T_s - T_u$. Moreover, we set $T_d = T_g$.

Let us denote with δ_{ij}^{RRH} the amount of RF capacity needed by a RRH RFB placed on node i to serve user $j \in \mathcal{U}$. This term can be expressed as in [10]:

$$\delta_{ij}^{RRH} = \left(\frac{B}{\sigma}\right)\left(\frac{T_{slot} - T_{pilot}}{T_{slot}}\right)\left(\frac{T_u}{T_s}\right)\log_2(1 + SIR_{ij}) \tag{5}$$

where B is the total system bandwidth, σ is the reuse factor, T_{slot} and T_{pilot} are the slot and the pilot duration, respectively, and SIR_{ij} is the Signal to Interference Ratio (SIR) experienced in the downlink between the RRH RFB located at node i and the user j. Specifically, we can express SIR_{ij} as:

$$SIR_{ij} = \frac{\beta_{ij}^2}{\sum_{p \neq i} \beta_{pj}^2} \tag{6}$$

The terms β_{ij} are defined as:

$$\beta_{ij} = \frac{z_{ij}}{s_{ij}^{\nu}} \tag{7}$$

[1] The evaluation of our system with more detailed radio link models is left for future work.

where z_{ij} is a log normal random variable, s_{ij} is the distance between the i-th node and the j-th user, and ν is the decay exponent. More in depth, $10\log_{10}(z_{ij})$ is zero mean Gaussian with standard deviation equal to $\omega_{(i)}^{shad}$.

Clearly, the user traffic has to be lower than the amount of capacity that is reserved by the RRH RFB to serve user j:

$$u_{ij}t_{ij} \leq \delta_{ij}^{RRH} \tag{8}$$

where u_{ij} is set to one if user j is connected to RRH RFB located at node i and t_{ij} is the amount of traffic to user j.

In addition, we assume that the total capacity of the RRH RFB consumed on the 5G node can be expressed as the sum of the RF capacities provided to users:

$$\delta_{i}^{RRH} = \sum_{j} \delta_{ij}^{RRH} \tag{9}$$

Finally, we assume that the total RF capacity to users has to not exceed the maximum capacity value R^{MAX} that can be handled by an RRH RFB:

$$\delta_{i}^{RRH} \leq R^{MAX} \tag{10}$$

C. BBU RFB Model

The BBU RFB module acts as an interface between the radio link managed by the RRH RFB and the video traffic provided by the MEC RFB. Specifically, a baseband traffic is exchanged between the RRH RFB and the BBU RFB. This amount of traffic requires the allocation of capacity resources on the node. More formally, the parameter δ_{i}^{BBU} (i.e., the amount of capacity consumed on the DHW to host baseband processing tasks of the BBU RFB) is computed from the model of [11]:

$$\delta_{i}^{BBU} = 2 \cdot S_R \cdot N_B \cdot A_i^G \cdot O_{CW} \cdot O_{LC} \tag{11}$$

where S_R is the sampling rate, N_B is the number of bits per sample, A_i^G is the number of antennas generating baseband traffic at site i, O_{CW} is the overhead introduced by the control words, and O_{LC} is the line coding overhead. Intuitively, the functions involving baseband operations require an high amount of capacity in the DHW part of the node.

Focusing then on the CPU processing tasks performed on the CHW part of the node, we assume that the CPU utilization of the BBU is composed of a static term that has to be counted if a BBU RFB is installed at node i, plus a dynamic term that scales with the amount of users traffic. More formally, we have:

$$C_i^{BBU} = C_i^{BS} + C_i^{BD} \sum_{j} u_{ij}t_{ij} \tag{12}$$

where C_i^{BS} is the static CPU utilization required by the BBU RFB and C_i^{BD} is a constant to transform the traffic from users into dynamic CPU utilization.

In addition, we have assumed that the memory utilization of the BBU RFB on the CHW scales with the number of users:

$$M_i^{BBU} = M_i^{BS} + M_i^{BD} \sum_{j} u_{ij} \tag{13}$$

where M_i^{BBU} is the memory utilization of the BBU, M_i^{BS} is the static memory utilization required by a BBU RFB, and M_i^{BD} is a constant to obtain the dynamic memory utilization, given the number of connected users.

D. MEC RFB Model

Finally, the MEC RFB module is responsible for providing the service to users. Similarly to the BBU case, the CPU and memory utilization of the MEC RFB on the CHW part of the node are defined as:

$$C_i^{MEC} = C_i^{MS} + C_i^{MD} \sum_{j} u_{ij}t_{ij} \tag{14}$$

$$M_i^{MEC} = M_i^{MS} + M_i^{MD} \sum_{j} u_{ij} \tag{15}$$

where C_i^{MS} and M_i^{MS} are static terms, while $C_i^{MD} \sum_j u_{ij}t_j$ and $M_i^{MD} \sum_j u_{ij}$ are dynamic ones.

E. Combined Model

We can infer some preliminary observations when the presented models are considered jointly together. The amount of served traffic t_{ij} from node i to user j depends on the capacity assigned to the radio link δ_{ij}^{RRH} by the RRH RFB, which, in turn, depends on: i) the user position, ii) the position of the 5G node where the RRH RFB is located, and, iii) the interference from the neighboring nodes. Moreover, the total amount of reserved capacity to users $\sum_j \delta_{ij}^{RRH}$ is bounded by the maximum capacity that can be handled by an RRH RFB R^{MAX}. In addition, the total amount of reserved capacity on the nodes (from both RRH and BBU RFBs) is bounded by the maximum amount of capacity B_i^{DHW} of the DHW part. Finally, the user traffic t_{ij} also influences the utilization of CPU and memory resources on the CHW part, which are also bounded by maximum values C_i^{CHW} and D_i^{CHW}. As a result, we can conclude that the users traffic heavily influences the management of the RFBs in the node.

Until now, we have focused on a single node and a single RFB chain. In a real network, however, multiple nodes, multiple chains, and multiple RFB types are deployed. Focusing on the RFB types, a macro cell may require an RRH RFB more demanding in terms of physical resources compared to an RRH RFB deployed for a small cell. Similarly, the baseband operations may require more resources for the RFBs serving macro cells, compared to the ones serving small cells. Therefore, it becomes of mandatory importance to develop a framework in order to optimize the RFBs management. To do that, in the next section we detail the problem formulation to manage the RFBs in a real network.

IV. Optimal Formulation

An informal description of the problem we tackle is the following:

- Given: the users positions in the considered scenario, the 5G nodes positions, the video requirements, the sets of RFBs, the RFBs features.

- Maximize: KPI.
- Subject to: RFBs placement constraints, 5G node capacity constraints, user coverage constraints and user data constraints.[2]

More formally, let us recall the set of nodes $\mathcal{N}$ and the set of users $\mathcal{U}$. In addition, we introduce the following sets: i) set of MEC RFBs types $\mathcal{K}^{MEC}$, ii) set of BBU RFBs types $\mathcal{K}^{BBU}$, iii) set of RRH RFBs types $\mathcal{K}^{RRH}$.

We first report the problem constraints and then we present the full problem formulations under different KPIs.

A. Problem Constraints

We first focus on the constraints related to RRH RFBs. Then, we detail the BBU and MEC RFBs constraints. Finally, we report the constraints of the 5G nodes.

1) RRH RFBs Constraints: First of all, we recall the binary variable u_{ij}, which takes value 1 if the user $j \in \mathcal{U}$ is served by node i, 0 otherwise. We then impose that each user has to be served by one 5G node:

$$\sum_{i} u_{ij} = 1 \quad \forall j \tag{16}$$

A user j can be served by node i only if one RRH RFB of type $k \in \mathcal{K}^{RRH}$ installed at node i is able to cover user j:

$$u_{ij} \leq \sum_{k} COV_{ijk} r_{ki} \quad \forall i, j \tag{17}$$

where COV_{ijk} is a binary input parameter taking value 1 if user j is covered by one RRH RFB of type $k \in \mathcal{K}^{RRH}$ installed on node i (0 otherwise), and r_{ki} is a binary variable taking value 1 if the RRH RFB of type k is installed on node i (0 otherwise). With this constraint, we impose also the fact that one RRH RFB has to be installed at node i if at least one user is connected to node i.

Moreover, the number of used RRH RFBs has to be lower than the total number of available RFBs of type k, denoted as N_k^{RRH}. More formally, we have:

$$\sum_{i} r_{ki} \leq N_k^{RRH} \quad \forall k \tag{18}$$

In addition, at most one RRH RFB is assigned to each node:

$$\sum_{k} r_{ki} \leq 1 \quad \forall i \tag{19}$$

Moreover, when an RRH RFB is installed at node i (i.e., $r_{ki} = 1$), the number of connected users is bounded by the maximum number of terminals for each RRH type k, which is denoted as U_k^{max}. More formally, the following constraint holds:

$$\sum_{j} u_{ij} \leq \sum_{k} U_k^{max} r_{ki} \quad \forall i \tag{20}$$

Each connected user will then receive an amount of RF capacity δ_{ikj}^{RRH}, which is computed from Eq.(5), by assuming that the RRH RFB of type k is installed on node i. The total capacity δ_{ik}^{RRH} provided by one RRH RFB of type k at node i is then computed as:

$$\delta_{ik}^{RRH} = \sum_{j} \delta_{ikj}^{RRH} r_{ki} u_{ij} \quad \forall i, k \tag{21}$$

δ_{ik}^{RRH} is then bounded by the maximum capacity that can be handled by the installed RRH RFB:

$$\delta_{ik}^{RRH} \leq R_k^{max} \quad \forall i, k \tag{22}$$

Moreover, the user traffic has to be lower than the RF capacity δ_{ikj}^{RRH}:[3]

$$t_{ij} r_{ki} \leq \delta_{ikj}^{RRH} \quad \forall i, j, k \tag{23}$$

where $t_{ij} \geq 0$ is a continuous variable representing the traffic between the node i and the user j. This variable has to be larger than zero only if the user j is assigned to the node i, as guaranteed by the following constraint:

$$t_{ij} \leq \mathcal{Q} u_{ij} \quad \forall i, j \tag{24}$$

where $\mathcal{Q}$ is a very large constant.

2) BBU and MEC RFBs Constraints: We initially focus on the BBU and MEC RFBs placement constraints. Specifically, an RFB chain composed by one RRH RFB, one BBU RFB and one MEC RFB has to be deployed in the network in order to serve the users connected to node i. Let us denote with b_{kip} a binary variable equal to 1 if one BBU RFB of type $k \in \mathcal{K}^{BBU}$ placed at node p is used to serve the RRH RFB at node i, 0 otherwise. If the node i has installed one RRH RFB of type w, then one BBU RFB has to serve it:

$$\sum_{k} \sum_{p} b_{kip} = \sum_{w} r_{wi} \quad \forall i \tag{25}$$

In addition, the number of used BBU RFBs is bounded by the number of available RFBs for each BBU type k, which is denoted as N_k^{BBU}:

$$\sum_{i} \sum_{p} b_{kip} \leq N_k^{BBU} \quad \forall k \tag{26}$$

Focusing on the MEC RFB case, we denote with m_{kip} a binary variable equal to 1 if one MEC RFB of type $k \in \mathcal{K}^{MEC}$ placed at node p is used to serve the users connected to the RRH RFB at node i, 0 otherwise. The MEC RFB constraint is then expressed as:

$$\sum_{k} \sum_{p} m_{kip} = \sum_{w} r_{wi} \quad \forall i \tag{27}$$

Clearly, the total number of used MEC RFBs is bounded by N_k^{MEC}, which is the number of available MEC RFBs of type k:

$$\sum_{i} \sum_{p} m_{kip} \leq N_k^{MEC} \quad \forall k \tag{28}$$

[2] The presented model can be extended to take into account also the capacity of links used to connect the nodes. This task will be done as future work.

[3] A parameter may be inserted here to take into account protocol overheads. We leave this aspect as future work.

Moreover, each RFB chain has to ensure compatibility between the RRH and BBU RFBs:

$$r_{ki} \sum_{p} b_{wip} \leq O_{kw} \quad \forall i, k, w \tag{29}$$

where O_{kw} is a binary input parameter, taking value 1 if a RRH RFB of type k and a BBU RFB of type w are compatible with each others, 0 otherwise. Intuitively, this constraint should prevent the connection of an RRH RFB designed for a macro cell with a BBU RFB designed for a small cell, which may otherwise introduce structural incompatibilities (e.g., not enough resources for the BBU RFB to serve the RRH RFB).

Finally, the total traffic to each user is then bounded by the HD video capacity provided by the MEC RFB:

$$t_{ij} \leq \sum_{p} \sum_{k} m_{kip} \delta_{k}^{MEC} \quad \forall i, j \tag{30}$$

3) 5G Nodes Constraints: We then focus on the constraints related to the 5G nodes. More in depth, the capacity used by RRH and BBU RFBs has to be lower that the one installed on the DHW part:

$$\sum_{k} r_{ki} \delta_{ik}^{RRH} + \sum_{w} \sum_{p} b_{wpi} \delta_{w}^{BBU} \leq B_{i}^{DHW} y_{i} \quad \forall i \tag{31}$$

where y_i is a binary variable taking value 1 if node i is used, 0 otherwise.

Moreover, the CPU utilization of the MEC RFBs installed at node i is computed as:

$$C_{i}^{MEC} = \sum_{k} \left[C_{ik}^{MS} c_{ik} + C_{ik}^{MD} \left(\sum_{p} m_{kpi} \sum_{j} t_{pj} \right) \right] \quad \forall i \tag{32}$$

where C_{ik}^{MS} and C_{ik}^{MD} are the static and dynamic terms introduced in the previous section to compute the CPU utilization, and c_{ik} is a binary variable, which takes the value one if at least one MEC RFB of type k is assigned to node i, 0 otherwise. We set c_{ik} with the following constraints:

$$\sum_{p} m_{kpi} \leq \mathcal{M} c_{ik} \quad \forall i, k \tag{33}$$

$$\sum_{p} m_{kpi} + e_{ik} \geq 1 \quad \forall i, k \tag{34}$$

$$e_{ik} + c_{ik} = 1 \quad \forall i, k \tag{35}$$

where $\mathcal{M}$ is a very large constant, and e_{ik} is a binary variable that is equal to 1 when no MEC RFB of type k is assigned to node i, 0 otherwise. The reason for introducing the last two constraints relies on the fact that we want to assure that c_{ik} is strictly set to zero when no MEC RFB of type k is installed in the node. In this way, in fact, the static amount of capacity C_{ik}^{MS} appearing in Eq. (32) is not counted.

Similarly, the amount of CPU consumed by BBU RFBs is computed as:

$$C_{i}^{BBU} = \sum_{k} \left[C_{ik}^{BS} d_{ik} + C_{ik}^{BD} \left(\sum_{p} b_{kpi} \sum_{j} t_{pj} \right) \right] \quad \forall i \tag{36}$$

where d_{ik} is a binary variable, which is computed in a similar way as in the MEC case:

$$\sum_{p} b_{kpi} \leq \mathcal{M} d_{ik} \quad \forall i, k \tag{37}$$

$$\sum_{p} b_{kpi} + f_{ik} \geq 1 \quad \forall i, k \tag{38}$$

$$f_{ik} + d_{ik} = 1 \quad \forall i, k \tag{39}$$

where f_{ik} is a binary variable that is equal to 1 if no BBU RFB of type k is assigned to the node i, 0 otherwise.

The total amount of used CPU resources on the CHW part is then bounded by the maximum number of CPU resources:

$$C_{i}^{MEC} + C_{i}^{BBU} \leq C_{i}^{CHW} y_{i} \quad \forall i \tag{40}$$

We then focus on the memory resources. Specifically, we express the amount of memory consumed by the MEC RFBs as:

$$M_{i}^{MEC} = \sum_{k} \left[M_{ik}^{MS} c_{ik} + M_{ik}^{MD} \left(\sum_{p} m_{kpi} \sum_{j} u_{pj} \right) \right] \quad \forall i \tag{41}$$

Moreover, we express the amount of memory consumed by the BBU RFBs as:

$$M_{i}^{BBU} = \sum_{k} \left[M_{ik}^{BS} d_{ik} + M_{ik}^{BD} \left(\sum_{p} b_{kpi} \sum_{j} u_{pj} \right) \right] \quad \forall i \tag{42}$$

The total amount of used memory resources is then bounded by the maximum number of memory resources:

$$M_{i}^{MEC} + M_{i}^{BBU} \leq M_{i}^{CHW} y_{i} \quad \forall i \tag{43}$$

B. Objective functions

Given the previous definitions of input parameters, variables and constraints we pursue the maximization of user throughput and the minimization of the number of used nodes as KPIs.

Maximization of user throughput. This KPI aims at maximizing the user performance. More formally, our problem is defined as:

$$\max \sum_{i,j} t_{ij} \tag{44}$$

subject to: (16)-(43); with control variables: u_{ij}, t_{ij}, r_{ki}, b_{kpi}, m_{kpi}.

Minimization of the number of used nodes. This objective aims to: i) limit the operating expenditures (OPEX) paid by operator (e.g., the node energy costs or the management ones), ii) efficiently exploit the nodes that are used. The following optimization problem is defined:

$$\min \sum_{i} y_{i} \tag{45}$$

subject to: (16)-(43); with control variables: u_{ij}, t_{ij}, y_i, r_{ki}, b_{kpi}, m_{kpi}.

In this latter case, t_{ij} is not constrained to a minimum value and it is not optimized. Therefore, an admissible solution is

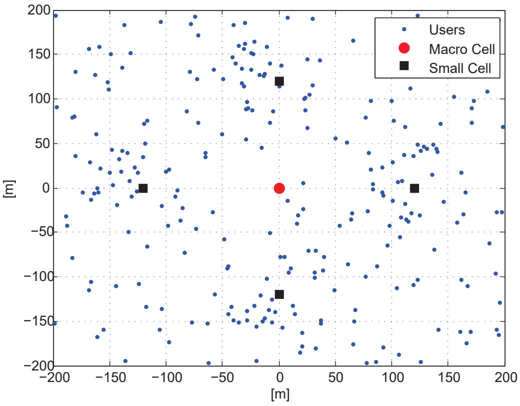


Fig. 4. Reference scenario with macro cell, small cells, and a realization of the users positions.

TABLE I
SUMMARY OF THE SYSTEM PARAMETERS

	Symbol	Value	[Source] / Appear in Eq.
MIMO	N_{OFDM}	7	[10] / Used to compute T_d in Eq. (4)
	τ	3	[10] / Eq. (4)
	T_u	66.7 μs	[10] / Eq. (4)
	T_c	500 μs	[10] / Used to compute T_d in Eq. (4)
	T_s	71.42 μs	[10] / Used to compute T_d in Eq. (4), Appears in Eq. (5)
	T_{slot}	500 μs	[10] / Eq. (5)
	B	20 MHz	[10] / Eq. (5)
	σ	1	[10] / Eq. (5)
	$\frac{(T_{slot}-T_{pilot})}{T_{slot}}$	3/7	[10] / Eq. (5)
	ν	3.8	[10] / Eq. (7)
	$\omega_{(i)}^{shad}$	8 dB	[10] / Used to compute the z_{ij} terms of Eq. (7)
BBU	S_R	30.72 MHz	[11] / Eq. (11)
	N_B	15	[11] / Eq. (11)
	O_{CW}	16/15	[11] / Eq. (11)
	O_{LC}	10/8	[11] / Eq. (11)

to set t_{ij} equal to zero for all the users. To avoid this issue, we rely on the ϵ-constrained method of [12] to force t_{ij} to be larger than zero. Specifically, we solve the problem with the objective of maximizing the users throughput, while we limit the number of used nodes by adding the following constraint to the problem:

$$\sum_i y_i \leq N_{used}^{max} \tag{46}$$

where N_{used}^{max} is the maximum number of used nodes, which is varied between 1 and $|\mathcal{N}|$. In this way, the optimal value of $\sum_i y_i$ is equal to the minimum value of N_{used}^{max} for which the problem satisfies the constraints (16)-(43),(46) while maximizing the users throughput.

TABLE II
RRH RFBS AND BBU RFBS PARAMETERS

	Parameter	Symbol	Value	
			RFB Type $k=1$	RFB Type $k=2$
RRH RFB	Maximum Number of Users	U_k^{max}	126	42
	Maximum Handled Capacity	R_k^{max}	29.96 [Gbps]	9.45 [Gbps]
	Number of RFBs	N_k^{RRH}	1	4
BBU RFB	Number of antennas generating traffic	A_i^G	126	42
	BBU capacity consumed on DHW	δ_k^{BBU}	156 [Gbps]	52 [Gbps]
	Number of RFBs	N_k^{BBU}	1	4

TABLE III
5G NODES PARAMETERS

	Parameter	Symbol	Value		
			Small Cell	Macro Cell	EPC
Capacity	DHW	B_i^{DHW}	122.91 [Gbps]	787.91 [Gbps]	727.99 [Gbps]
	CHW (CPU/Mem.)	C_i^{CHW} M_i^{CHW}	2 [units]	4 [units]	4 [units]
MEC/BBU Util.	CPU Static	C_{ik}^{BS} C_{ik}^{MS}	0.5 [units]	0.5 [units]	0.5 [units]
	CPU Dyn.	C_{ik}^{BD} C_{ik}^{MD}	$5.28 \cdot 10^{-5}$ [1/Mbps]	$7.37 \cdot 10^{-6}$ [1/Mbps]	$7.37 \cdot 10^{-6}$ [1/Mbps]
	Mem. Static	M_{ik}^{BS} M_{ik}^{MS}	0.5 [units]	0.5 [units]	0.5 [units]
	Mem. Dyn.	M_{ik}^{BD} M_{ik}^{MD}	0.0116 [units]	0.0019 [units]	0.0019 [units]

V. SCENARIO DESCRIPTION

We consider a scenario composed of one macro cell, four small cells, and 260 users requesting 5G services. We assume that this scenario is representative for the peak traffic condition. Fig. 4 reports the cells and the user positions. More in depth, the macro cell is placed in the center of the service area. Each small cell is placed at a distance of 120 [m] far from the macro cell. We assume that small cells may interfere with each others, while the central macro cell may interfere with a set of neighboring macro cells, placed at the corners of a square centered by the considered macro cell, with an edge equal to 1000 [m]. Focusing on users, 70% of them are randomly deployed over the whole service area, while 30% are generated in a circle of radius equal to 50 [m] centered in each small cell (thus justifying the small cell deployment).

Focusing on the RFBs, we assume a total of 5 RRH RFBs, 5 BBU RFBs, and 5 MEC RFBs. In addition, we assume two types of RRH RFBs, two types of BBU RFBs, and one type of MEC RFB. The intuition of having two types of RRH RFBs and BBU RFBs relies on the fact that the traffic handled by the macro cell node is in general higher than the one of a small cell. Therefore, the resource requirements of the associated RFBs may be different, resulting in two different RFB types.

Tab. I reports the settings of the MIMO and BBU parameters, which relies on the works [10], [11]. In addition, the setting of the RRH RFBs and BBU RFBs parameters is reported in Tab. II, respectively. More in depth, U_k^{max} is computed from Eq. (4), by assuming that the RRH RFB of the macro cell is composed of 3 sectors. In addition, R_k^{max} is computed in the following way (for each RRH type): i) each user is assigned to the cell i^* of type k^* providing the highest SIR; ii) each user j receives the maximum capacity value $\delta_{i^*k^*j}^{RRH}$ from the associated cell; and, iii) the total capacity for each RRH RFB is then computed as the maximum capacity over the other nodes with the same type k^*. Focusing then on the BBU RFBs, the

BBU parameters of Tab. I are plugged into Eq. (11), in order to get the total BBU RFB capacity consumed on the DHW part δ_k^{BBU} (reported in Tab. II). Not surprisingly, each BBU RFB requires a substantial higher amount of capacity w.r.t. the capacity managed by an RRH RFB. Moreover, we assume the following values for the compatibility between modules: $O_{11} = 1$ and $O_{22} = 1$. Finally, we set the total capacity of the MEC RFB as: δ_k^{MEC}=29.96 [Gbps], i.e., the maximum capacity of a MEC RFB is equal to the maximum handled capacity R_k^{max} by an RRH RFB of a macro cell.

Once the RFBs capacities have been expressed, the next step is to properly set up the nodes resources. Specifically, we adopt the following assumptions: i) the network has to satisfy the amount of traffic generated by users with the RFBs deployed in the nodes; ii) the resources of each small cell node are set to host at least one RRH RFB and one BBU RFB for the DHW, and one BBU RFB and one MEC RFB in the CHW; iii) the macro cell node and the EPC node are designed to pool the BBU and MEC RFBs from the small cells; and, iv) an amount of spare resources is always reserved in each node (i.e., to cope with future traffic increases). Tab. III reports the parameters for the CHW and DHW parts of the 5G nodes. Specifically, we express B_i^{DHW} in terms of [Gbps], while we decided to express C_i^{CHW} and M_i^{CHW} in terms of [units]. The reason for this choice is that B_i^{DHW} is directly related to the bandwidth consumed by the RFB on the DHW part of the node, while C_i^{CHW} and M_i^{CHW} depend on the CPU and memory utilizations. The effective definition of C_i^{CHW} and M_i^{CHW} in terms of measurement units will be done as future work.[4] In addition, the table reports also the parameter settings for the static and the dynamic utilization. Specifically, in order to introduce a gain when the RFBs are pooled together in the same node, we have assumed a static utilization of 0.5 [units] for both CPU and memories, i.e., there is a high cost in deploying a single RFB on the node. Then, this cost is shared as long as other RFBs of the same type are placed on the same node. The dynamic utilization, which represents the slope of the utilization functions in Eq. (12)-(15), is designed to have an utilization of resources lower than the maximum one (e.g., when 4 BBU RFBs of type 2, 1 BBU RFB of type 1 and 5 MEC RFBs are installed on the macro cell or the EPC nodes).

VI. Performance Evaluation

We solve the proposed optimization problem over the considered scenario on a high performance computing cluster, composed of four nodes, each of them with 32 cores and 64 GB of RAM, for a total computing power of around 1.5 TeraFlops/s.[5] In addition to the peak traffic condition, we take into account the case in which only 10% of users generate traffic, which is referred as "off peak" from now on.[6] Our goal is in fact to assess the performance of the considered NFV architecture under different traffic conditions.

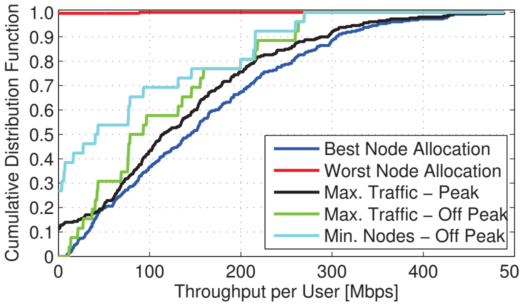
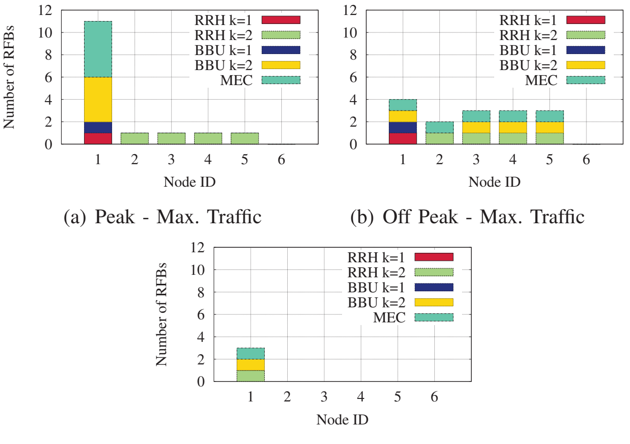


Fig. 5. CDFs of traffic assigned to users.



(a) Peak - Max. Traffic

(b) Off Peak - Max. Traffic

(c) Off Peak - Min. Nodes

Fig. 6. RFBs placement for peak and off peak traffic (ID 1 = Macro cell, ID 2-5 = Small cell, ID 6 = EPC Node).

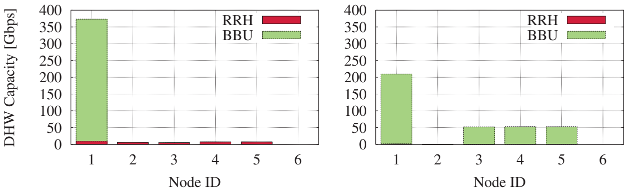
We initially take into account the amount of traffic t_{ij} served to each user. More in depth, Fig. 5 reports the Cumulative Distribution Function (CDF) of the user traffic for the two traffic conditions. The figure reports also the CDFs for the following cases: i) best node allocation policy, i.e., $t_{ij} = \max_i \delta_{ikj}^{RRH}$; and, ii) worst node allocation policy, i.e., $t_{ij} = \min_i \delta_{ikj}^{RRH}$. Interestingly, the traffic served to users during the peak traffic condition is close to one achieved by the best node allocation, with an average traffic of more than 100 [Mbps] per user. However, a small subset of users (around 10%) is experiencing very low traffic (i.e., close to 0). By further investigating this issue, we have found that such users are close to the edges of the macro cell, i.e., they are the ones experiencing the worst channels conditions. To overcome this issue, these users could be potentially covered also by the neighboring macro cells in a real environment. In addition, Fig. 5 reports also the CDF when the off peak traffic condition is considered. In this case, the average traffic per user is still higher than 80 [Mbps], while no user exhibits extremely low traffic conditions.[7] Finally, the figure reports the CDF of the

[4] Intuitively, C_i^{CHW} may represent the number of installed CPU cores, while M_i^{CHW} may denote the amount of RAM used.

[5] We have linearized the nonlinear constraints of the optimization problem. The linearization is not included in this work due to the lack of space.

[6] The off peak users are randomly selected from the peak ones.

[7] We have manually verified that no user in the selected off peak set of users is experiencing bad channel conditions. The repetition of this experiment with different subsets of randomly selected users is left for future work.



(a) Peak (b) Off Peak

Fig. 7. Used DHW capacity for peak and off peak traffic (Max. Traffic)

user throughput for the off peak traffic condition when the minimization of used nodes is pursued (i.e., by adopting the ϵ-constrained method of Sec. IV-B).[8] In this case, the users are connected to a single node in the network, i.e., the macro cell. Clearly, this choice has an impact on the throughput, which tends to be decreased (i.e., the average throughput is less than 50 [Mbps]).

Fig. 6 reports the RFBs placement over the set of nodes. Focusing on the peak traffic condition and the maximization of the user throughput (Fig. 6(a)), all the RRH RFBs are exploited, in order to maximize the performance to users. In addition, the BBU and MEC RFBs are all located on the macro cell node. Then, Fig. 6(b) reports the off peak traffic. In this case, the BBU and MEC RFBs tend to be spread over the nodes. This is due to the fact that the number of used nodes is not taken into account when the objective function is solely the maximization of the users traffic. As a result, all the nodes may be potentially exploited even if the number of users is low. Focusing on the same traffic condition and on the minimization of the number of used nodes (Fig. 6(c)), the macro cell node is hosting one RRH RFB of type 2, one BBU RFB of type 2 and one MEC RFB.

In the next part, we consider the utilization of nodes resources consumed by the RFBs when the maximization of the user throughput is pursued. Focusing on the DHW part, Fig. 7(a) reports the amount of used capacity for the peak traffic condition. As expected, the largest amount of capacity is consumed by the BBU RFBs, while the RRH RFBs marginally impact the overall capacity. This is due to the fact that BBU RFBs perform the baseband operations, which are pretty intensive on the computing resources of the node. Moreover, the total capacity consumed on the macro cell node is more than 350 [Gbps], i.e., close to 50% of the installed capacity. On the contrary, the small cell nodes are lightly utilized. Moreover, Fig. 7(b) presents the results for the off peak traffic condition. Eventually, the amount of capacity consumed on the small cell nodes tends to promptly increase, as the BBU RFBs are deployed also on most of them. Nevertheless, the amount of capacity used on the macro cell node is still higher than 200 [Gbps] (i.e., more than 20%).

VII. Conclusions and Future Works

We have considered the RFBs management over a NFV-based 5G network, as an outcome of the SUPERFLUIDITY project. After modelling the different RFBs types and the 5G nodes hosting them, we have optimally formulated the problem of dynamically managing the RFBs under the following objective functions: i) maximization of the user traffic; ii) minimization of the number of used nodes. We have solved the problem over a simple, yet representative, scenario. Our results show that: i) users are able to achieve very good throughput in the downlink direction; ii) the RFBs placement is impacted by the number of users and the considered strategy; iii) the BBU RFBs of the same type and the MEC RFBs may be efficiently pooled on the same node to better exploit the CHW and DHW resources. As next step, we plan to consider the impact on the uplink direction, more realistic channel models for the radio link, and a more detailed simulation of the scenarios. In addition, we will solve the proposed problem on a metropolitan scenario composed of a larger number of 5G nodes, and we will consider the impact of RFBs on the bandwidth of the physical links connecting the 5G nodes. Finally, we will consider more complex models for RFB chaining, as well as differentiated services to users.

Acknowledgments

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

References

[1] G. P. Fettweis, "The tactile internet: applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, 2014.

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] R. Jain and S. Paul, "Network virtualization and software defined networking for cloud computing: a survey," *Communications Magazine, IEEE*, vol. 51, no. 11, pp. 24–31, 2013.

[4] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: A survey," *Future Generation Computer Systems*, vol. 29, no. 1, pp. 84–106, 2013.

[5] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks: a technology overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[6] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[7] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[8] 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.

[9] "ETSI GS NFV 002: Network Functions Virtualisation (NFV); Architectural Framework, V 1.2. 1," *ETSI, December*, 2014.

[10] 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.

[11] M. Fiorani, B. Skubic, J. Mårtensson, L. Valcarenghi, P. Castoldi, L. Wosinska, and P. Monti, "On the design of 5G transport networks," *Photonic Network Communications*, vol. 30, no. 3, pp. 403–415, 2015.

[12] G. Mavrotas, "Effective implementation of the ε-constraint method in multi-objective mathematical programming problems," *Applied mathematics and computation*, vol. 213, no. 2, pp. 455–465, 2009.

[8] The optimization results with the minimization of the number of used nodes and the peak traffic condition are exactly the same obtained when the goal of the problem is solely the maximization of the user throughput.