Optimal Management of Reusable Functional Blocks in 5G Superfluid Networks

Article *in* International Journal of Network Management · June 2018

DOI: 10.1002/nem.2045

6 authors, including:

Luca Chiaraviglio
University of Rome Tor Vergata
143 PUBLICATIONS **3,235** CITATIONS
SEE PROFILE

Lavinia Amorosi
Sapienza University of Rome
14 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Mohammad Shojafar
Ryerson University
99 PUBLICATIONS **1,878** CITATIONS
SEE PROFILE

Stefano Salsano
University of Rome Tor Vergata
182 PUBLICATIONS **2,114** CITATIONS
SEE PROFILE

Optimal Management of Reusable Functional Blocks in 5G Superfluid Networks

Luca Chiaraviglio,[1,2] Lavinia Amorosi,[3] Nicola Blefari-Melazzi,[1,2] Paolo Dell'Olmo,[3] Mohammad Shojafar,[1] Stefano Salsano[1,2]

1) Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy, email luca.chiaraviglio@gmail.com

2) Department of Electronic Engineering, University of Rome Tor Vergata, email {name.surname}@uniroma2.it

3) Department of Statistical Sciences, University of Rome Sapienza, Rome, Italy, email {name.surname}@uniroma1.it

***Abstract*—We consider the problem of managing a 5G network composed of virtualized entities, called Reusable Functional Blocks (RFBs), as proposed by the Horizon 2020 SUPERFLUIDITY project. The RFBs are used to decompose network functions and services, and are deployed on top of physical nodes, in order to realize the 5G functionalities. After formally modelling the RFBs in a 5G network, as well as the physical nodes hosting them, we formulate the problem of managing the 5G network through the RFBs, in order to satisfy different Key Performance Indicators (KPIs) to users. In particular, we focus either on the maximization of the amount of downlink throughput sent to users, or on the minimization of the number of powered-on physical nodes. We then consider different scenarios to evaluate the proposed formulations. Our results show that, when an RFB-based approach is put into place, a high level of flexibility and dynamicity is achieved. In particular, the RFBs can be shared, moved, and rearranged based on the network conditions. As a result, the downlink throughput can be extremely high, i.e., more than 150 [Mbps] per user on average when the throughput maximization is pursued, and more than 100 [Mbps] on average when the goal is the minimization of the number of powered-on physical nodes.**

I. Introduction

The forthcoming 5G technology will allow the deployment of a variety of new applications, such as the streaming of Ultra High Definition (UHD) videos, the introduction of the tactile Internet [2], the increase in the number of connected devices through the Internet of Things (IoT) paradigm, and the possibility to transfer information with extremely low delay. Moreover, the 5G technology will be able to sustain the increase in the number of connected users, especially in very crowded environments, such as stadiums, airports, train stations and shopping malls. In order to meet these ambitious goals, the design from the scratch of the entire network architecture will be a mandatory step, aiming at the deployment of a fully converged, flexible, and extremely high performance network. In this context, 5G is going to extensively exploit new paradigms, such as 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]. In addition, several initiatives are devoted to the design of 5G networks [7], which are expected to turn into reality by 2020.

Among the different projects currently investigating 5G architectures, it is worthwhile to mention SUPERFLUIDITY [8], which is funded by the EU through the H2020 program. The goal of SUPERFLUIDITY is to design a new 5G network architecture, which ensures the required levels of flexibility, agility, portability and high performance. In a nutshell, SUPERFLUIDITY aims to achieve a *superfluid* state of the network, which is the ability to instantiate services on-the-fly, run them anywhere in the network (core, aggregation, edge) and shift them transparently to different locations. The core of the project is the definition of the concept of Reusable Functional Block (RFB), which is a virtualized entity, used to decompose network functions and services, and it is deployed on top of a physical node. The deployment of the 5G network through the RFBs have notable features, including: i) the possibility to build chain of RFBs, in order to implement more complex functionalities and to provide different services to users; ii) the independence of the RFBs from a specific platform, i.e., RFBs can be realized via software functions, and they can be run on several HardWare (HW) architectures; and, iii) the introduction of high levels of flexibility and performance, thanks to the fact that the RFBs can be deployed where and when they are really needed. 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 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 the RFBs in a 5G superfluid network? How to model the RFBs types and the physical HW hosting them? How to optimally map the RFBs on the underlying HW under different Key Performance Indicators (KPIs)? What is the impact of the scenario on the obtained results? The answer to these questions is the goal of the paper. Specifically, we con-

A preliminary version of this work has been presented at the IEEE NetSoft 2017 conference [1]. Corresponding author: Luca Chiaraviglio, Department of Eletronic Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy.

sider a cloud-based 5G architecture to model the RFBs types and the physical HW features. We then optimally formulate the problem of managing a set of RFBs in order to serve the users of a 5G network when a HD video distribution service is offered. In our analysis, we target either the maximization of the downlink throughput to users, or the minimization of the number of physical HW devices powered on. Our results, obtained over a set of representative case studies, show that, when the maximization of the throughput to users is pursued, each user can receive more than 150 [Mbps] of traffic on average. On the other hand, when the goal is the minimization of the number of physical nodes powered on, a throughput larger than 100 [Mbps] is ensured to each user on average.

Even though the results presented in this paper are promising, we point out that the considered 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 decomposition of RFBs into smaller RFBs, the mapping of RFBs to different SW environments, as well as the introduction of recursive interactions between the RFBs, will be interesting branches of future research.

The rest of the paper is organized as follow. The related works are reviewed in Sec. II. The RFB concept and its novelty compared to VNF are detailed in Sec. III. Sec. IV describes the considered 5G network architecture. Sec. V details the considered RFBs and the physical HW models. The problem formulations under different KPIs are detailed in Sec. VI. The 5G scenarios under investigation are described in Sec. VII. Sec. VIII details the obtained results. Finally, Sec. IX reports the conclusions and possible future works.

II. Related work

The explosive growth in traffic volumes, the huge increase in the number of connected wireless devices, and the wide range of QoS requirements of 5G devices impose to face critical research challenges. Therefore, the design of 5G networks includes a wide range of technological advancements from transmission techniques (e.g. MU-MIMO) to network architectures (C-RAN, NFV, MEC). As a result, we divide the related work in the following categories: transmission technologies, edge-based architectures, management of 5G services/functionalities, virtualization of network functions, and related EU initiatives for comprehensive 5G architectures.

A. Transmission Technologies

An introduction to the research challenges of 5G networks, mainly focusing on the transmission technologies, can be found in [10]. The work in [11] tackles the problem of optimal allocation of radio resources among different tiers of radio transmissions (e.g. macro cells and micro cells). In [12] the authors focus on similar radio level issues in the 5G wireless backhaul networks. Recent advances in radio transmission technologies include much greater spectrum with mmWave frequency spectrum bands [13], highly directional massive beamforming antennas for mobile and stand-alone devices [14], full-duplexing communications (FDCs) [15], and higher aggregate capacity thanks to heterogeneous networks [16]. Although all these solutions are beneficial to the deployment of mobile networks, they lack coverage of the MEC technologies.

B. Edge-based Architectures

The need of reducing the latency in accessing the services leads to rethink the traditional topology of the service infrastructure, by moving the applications to the network edge, closer to end user devices. The concept of Cloudlets (locally distributed cloud computing environments) is discussed in [17] and [18]. The fog computing approach is proposed in [19]. Following this trend towards *Edge Computing*, the MEC Industry Specification Group (ISG) was created in December 2014 [20], [21]. A tutorial on MEC technology can be found in [22]. The original meaning of MEC was *Mobile Edge Computing*, then it was recently re-branded by ETSI as *Multi-access Edge Computing* [23] to better reflect the scenarios with non-cellular access networks in 5G. A key aspect of this Edge Computing revolution is the possibility to achieve and exploit the integration among radio level, network level and services/application level technologies. In [24], a MEC platform is introduced in which the C-RAN computing platform is fully integrated with MIMO technology for the mobile user and Base Stations. Such platform facilitates accessing data that has not been exploited to date (e.g., cell congestion, user locations, and movement direction) to build 5G services and applications.

Considering the offloading of processing from the mobile device to the edge computing platform, in [25] the authors provide a mathematical formulation of the computation offloading problem aimed at optimizing the communication and computation resources jointly together, posing a strict attention to latency and energy constraints. Additionally, the authors in [26] highlighted a robust solution for the offloading problem tailored to MEC scenarios and optimize the overall energy consumption of the engaged components at the mobile terminal sides, under transmit power and latency constraints. However, one major limitation of their method is the limited flexibility of the considered MEC solution. In contrast to them, our work is more focused to the dynamic management of the resources, as well as to the possibility to host these resources on a heterogeneous set of physical nodes.

C. Management of 5G Services/Functionalities

One of the key aspect arising during the operation of a 5G network is the management of 5G services and virtualized functionalities over the physical infrastructure. In fact, the network-aware combination of SDN and NFV results in generic HW boxes and SW components that have to be managed across network segments. In this context, solutions based on softwarization allow advanced configuration and customization of the network functions. Recently, the authors in [27] considered this problem and developed a self-healing framework for an SDN-based 5G network. Their framework manages the availability of the services, network functions, and engaging resources over SDN-based networks. Moreover, in [28] the authors deal with the same problem by proposing a high level concept based on dynamic SW

module placements in a cloud-based infrastructure to support the 5G network and services requirements. In [29] the authors discusses opportunities of the NFV approaches in reducing the CAPEX/OPEX costs and the challenges of developing an integrated management framework for an NFV-based 5G network infrastructure. In [30] the authors provide a techno-economic analysis for the introduction of softwarization technologies in the 5G network architecture, with a cost model to estimate the CAPEX and OPEX of the proposed architecture. In [31] the authors proposed two scheduling solutions to manage the Remote Radio Heads (RRH) unit and control inter-cell interference (ICI). They assumed that small cell coverage area can be dynamically divided in different numbers of sectors to take into account traffic loads and interference suffered from interior and cell edge users. In contrast to them, our work exploits the MU-MIMO technology. In addition, we take advantage of the softwarization paradigm to place the radio resources in a flexible way, and move them across the physical nodes when needed.

D. Virtualization of Network Functions

A first of set of works is devoted to the investigation of the VNF placement problem (see e.g., [32], [33], [34], [35], [36]). In particular, [32] targets the placement of the network functions, as well as their chaining, by taking into account the amount of available resources and the requirements of the functions. Moreover, the possible trade-offs between different optimization objectives are investigated. However, the work in [32] is only focused on the transport part of the Internet Service Provider (ISP), thus completely neglecting the radio access network. In a similar way, [33] is focused on a hybrid scenario where the services are provided either by dedicated physical hardware or by virtualized service instances. The optimal placement of VNFs and the optimal assignment of demands to the VNF chains are investigated by [34]. The proposed solution requires as input the set of sources and destinations, as well as the set of traversed VNFs for each demand. Similarly to [33], the radio access network is not taken into account at all in [33], [34]. A first step towards a more global approach is proposed in [35], where authors assume that VNFs can be placed in commodity HW installed in Data Centers (DCs), or on top of routers/switches. Their goal is to reduce the total consumption of the network while meeting the service requirements for all the traffic flows. Also in this case the radio access network is not modeled at all. Finally, the authors of [36] extends the VNF placement problem to Content Delivery Networks (CDNs), but still not considering the radio part of the network.

A second taxonomy includes the works investigating the virtualization of the radio access networks (see e.g., [37], [38], [39], [40], [41], [42]). More in depth, the authors of [37] advocate the need for a flexible network design for next generation networks. However, the work is tailored to the architectural level, and no optimization model is proposed. The authors of [38] focus on mobile packet core network architectures based on SDN/NFV. In particular, the optimization of the functions placement, the resource allocation, the management and the orchestration are listed among the different challenges and issues that need to be faced for this type of architectures. Moreover, in [39] the authors targets the virtualization of a radio access network, either based on Virtual Machines (VMs) and Docker containers. Results demonstrate the superiority of the Docker technology compared to a classical VM-based one. However, also in this case the model formalization is not taken into account.

Eventually, the authors in [40] detail an optimization problem for the placement of BBUs in a Cloud Radio Access Network (C-RAN) architecture running over a Wavelength Division Multiplexing (WDM) aggregation network. However, the MEC placement, as well as the modeling of the communication channel between each user and the serving RRH, are not considered. The authors of [41] focus on the problem of joint optimization of cloud and edge processing in fog radio access networks. In particular, the authors assume that the RRHs may be equipped with local caches, containing frequently requested content and baseband processing capabilities. The problem of maximizing the delivery rate is formulated under the capacity of the fronthaul and the power constraints of the RRHs. The presented problem is focused on coding, based on the exploitation of different modes available on the fronthaul links. However, the placement of the functions in a general network topology is not considered at all. Eventually, the work in [42] is devoted on the design phase of a 5G RFB-based radio network. More in depth, the goal is to decide where and which nodes to install, based on their installation costs, as well as selecting the available RFBs. On the contrary, in this work we focus on the management of the 5G network: given the set of installed nodes and the set of available RFBs, our goal is to efficiently manage the RFB resources in order to target a given objective (i.e., maximization of the user throughput, or minimization of the number of used nodes). Consequently, the two works investigate complementary problems: the output of the design phase is used as input for the management one, in order to efficiently exploit the available resources.

E. Related EU Initiatives

Finally, different EU HORIZON 2020 projects have provided comprehensive architectures to manage network components and connectivity in dense networks (see, e.g., [43], [44], [45]). Among them, the SELFNET project [45] aims to generate a significant impact on the development of 5G, mainly in societal, operational, and innovation levels. The project aims to define a new management framework upon the software-defined and virtualized network concepts [46]. The SELFNET framework is basically different from the one considered in this work, due to the following reasons: (i) although this method is effective in app orchestration, it lacks controlling the user traffic signals, the data requests and the resource management for the heterogeneous devices; (ii) the virtualization of Base Band Units (BBUs) and MEC resources are not the main focus of the project. On the other hand, in our proposed approach we are more focused on the management of RRH, BBU and MEC components. In addition, we face the problem on how the virtual resources can be mapped on the physical ones in a flexible way through the RFB concept.

III. Reusable Functional Blocks in 5G Networks: Concept and Novelty

A key aspect of the SUPERFLUIDITY architecture is that the needed components are built as a composition of elementary building blocks, which are completely softwarized. The management of the building blocks over the physical resources can be very dynamic, in order to instantiate the softwarized components in real time. The composition of such components is based on the concept of RFB, which is a logical entity performing a set of functionalities. Each RFB is then denoted by a set of input and output ports. Moreover, an RFB can also include state information, so that the actual output depends on the input and the state saved inside the RFB. A composition of multiple RFBs generates a service or a more complex RFB (hence the term "Reusable"). Since the concept of RFB is not constrained to a given operating system, platform-agnostic languages and tools are needed to describe the interactions and connections between the RFBs and the execution environments on which the RFB can be deployed. We refer the reader to [8] for a detailed overview of these features.

Focusing then on the practical modeling of the RFBs, these components require resources in terms of storage and/or processing. In addition, they may be described in terms of maximum achievable performance, e.g., maximum number of packets that can be processed per second. Finally, the input and the output are defined in order to properly connect the RFBs together.

It is worth highlighting the main differences between the VNF concept considered in the ETSI NFV architecture [9] (illustrated in particular in [47]) and the proposed RFB concept. According to [47] a VNF is made up by one or more *VNF Components* (VNFCs) that can be deployed in the NFVI infrastructure. A VNFC can be implemented as a VM running on a hypervisor or with OS container technology (e.g. Docker). Therefore, VNFCs cannot be further decomposed in other VNFs and their execution environment is restricted to hypervisors and OS containers. One of the key benefit introduced by the RFB concept is that the composition of a service through a set of RFBs can be applied in very different and heterogeneous environments (including of course the hypervisors and OS containers representing the traditional NFV Infrastructure). In particular, an RFB can be implemented as a lightweight Unikernel VM running on a custom hypervisor. Moreover, other possible implementation of RFBs can be modules and components of special purpose execution environments, like extended finite state machines based on OpenFlow for packet processing [48], software routers [49], or radio signal processing chains [50]. In this way, the management of the network through a set of RFBs introduces the levels of flexibility and agility which are required for a full exploitation of the predicted 5G services [8].

IV. Architecture Description

The 5G network model considered in this work is composed of a set of physical 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 physical 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 one). For simplicity, the EPC elements are collapsed in a single site in our model.

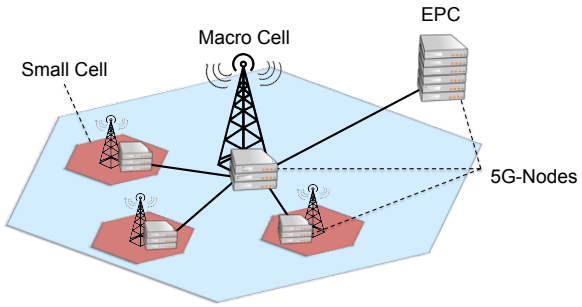


Fig. 1. Physical system infrastructure.

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* in the rest) 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;
- Remote 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 one. 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.

RRH RFB. This module performs physical layer operations. Specifically, the RRH module handles a set of Radio

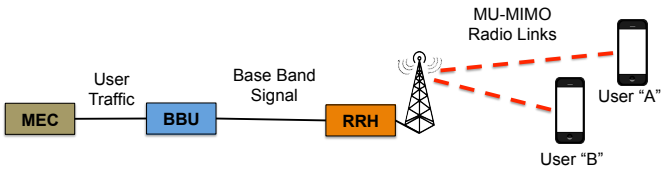


Fig. 2. RFBs relationship and exchanged information.

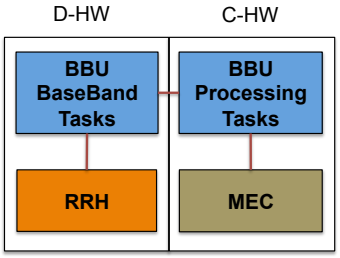


Fig. 3. 5G-Node architecture. The Commodity Hardware (CHW) hosts MEC RFBs and BBU processing tasks. The Dedicated Hardware (DHW) hosts BBU baseband tasks and RRH RFBs.

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 one).

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 an 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 5G nodes powered on.

V. 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) part and a Commodity HardWare (CHW) one. 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 Radio Frequency (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 one 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 one, 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 [51], we assume that the number of installed antennas is larger than the number of users served by the cell (either a small cell or a macro one). In this way, we can rely on [51] 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]

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 site and small cell one for the moment). U^{max} is bounded by the reverse link constraint of [51]:

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

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 an RRH RFB placed on node i to serve user $j \in \mathcal{U}$. This term can be expressed as in [51]:

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

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:

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

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:

$$\delta_i^{RRH} = \sum_j \delta_{ij}^{RRH} \tag{9}$$

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

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 [52]:

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

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. Interactions among the Models

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 R^{MAX} that can be handled by an RRH RFB. In addition, the total amount of reserved capacity on the nodes (for 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 RFBs 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.

VI. Problem 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:

- set of MEC RFBs types $\mathcal{K}^{MEC}$,
- set of BBU RFBs types $\mathcal{K}^{BBU}$,
- set of RRH RFBs types $\mathcal{K}^{RRH}$.

We first report the problem constraints, then we present the linearization of the non-linear constraints, and finally we detail the formulations under different KPIs.

A. Problem Constraints

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

[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.

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:

$$\sum_{i \in \mathcal{N}} u_{ij} = 1 \quad \forall j \in \mathcal{U} \tag{16}$$

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:

$$u_{ij} \leq \sum_{k \in \mathcal{K}^{RRH}: i \in RRH_k} COV_{ijk} r_{ki} \quad \forall i \in \mathcal{N}, j \in \mathcal{U} \tag{17}$$

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), r_{ki} is a binary variable taking value 1 if the RRH RFB of type k is installed on node i (0 otherwise) and RRH_k is the set of 5G nodes where an RRH of type k can be placed. 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:

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

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

$$\sum_{k \in \mathcal{K}^{RRH}} r_{ki} \leq 1 \quad \forall i \in \mathcal{N} \tag{19}$$

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:

$$\sum_{j \in \mathcal{U}} u_{ij} \leq \sum_{k \in \mathcal{K}^{RRH}} U_k^{max} r_{ki} \quad \forall i \in \mathcal{N} \tag{20}$$

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:

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

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

$$\delta_{ik}^{RRH} \leq R_k^{max} \quad i \in \mathcal{N}, k \in \mathcal{K}^{RRH} \tag{22}$$

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

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

$$t_{ij} r_{ki} \leq \delta^{RRH}_{ikj} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{23}$$

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:

$$t_{ij} \leq \mathcal{M} u_{ij} \quad \forall i \in \mathcal{N}, j \in \mathcal{U} \tag{24}$$

where $\mathcal{M}$ 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.[4] 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:

$$\sum_{k \in \mathcal{K}^{BBU}} \sum_{p \in \mathcal{N}} b_{kip} = \sum_{w \in \mathcal{K}^{RRH}} r_{wi} \quad \forall i \in \mathcal{N} \tag{25}$$

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^{BBU}_k:

$$\sum_{i \in \mathcal{N}} \sum_{p \in \mathcal{N}} b_{kip} \leq N^{BBU}_k \quad \forall k \in \mathcal{K}^{BBU} \tag{26}$$

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:

$$\sum_{k \in \mathcal{K}^{MEC}} \sum_{p \in \mathcal{N}} m_{kip} = \sum_{w \in \mathcal{K}^{RRH}} r_{wi} \quad \forall i \in \mathcal{N} \tag{27}$$

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

$$\sum_{i \in \mathcal{N}} \sum_{p \in \mathcal{N}} m_{kip} \leq N^{MEC}_k \quad \forall k \in \mathcal{K}^{MEC} \tag{28}$$

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

$$r_{ki} \sum_{p \in \mathcal{N}} b_{wip} \leq O_{kw} \quad \forall i \in \mathcal{N}, k \in \mathcal{K}^{RRH}, w \in \mathcal{K}^{BBU} \tag{29}$$

where O_{kw} is a binary input parameter, taking value 1 if an RRH RFB of type k and a BBU RFB of type w are compatible with each other, 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 one). Finally, the total traffic to each user is then bounded by the HD video capacity provided by the MEC RFB:

$$t_{ij} \leq \sum_{p \in \mathcal{N}} \sum_{k \in \mathcal{K}^{MEC}} m_{kip} \delta^{MEC}_k \quad \forall i \in \mathcal{N}, j \in \mathcal{U} \tag{30}$$

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:

$$\sum_{k \in \mathcal{K}^{RRH}} \delta^{RRH}_{ik} + \sum_{w \in \mathcal{K}^{BBU}} \sum_{p \in \mathcal{N}} b_{wpi} \delta^{BBU}_w \leq B^{DHW}_i y_i \quad \forall i \in \mathcal{N} \tag{31}$$

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:

$$\begin{aligned} C^{MEC}_i &= \textstyle\sum_{k \in \mathcal{K}^{MEC}} \Big[C^{MS}_{ik} c_{ik} + \\ &+ C^{MD}_{ik} \left(\textstyle\sum_{p \in \mathcal{N}} m_{kpi} \sum_{j \in \mathcal{U}} t_{pj} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{32}$$

where C^{MS}_{ik} and C^{MD}_{ik} 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:

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

$$\sum_{p \in \mathcal{N}} m_{kpi} + e_{ik} \geq 1 \quad \forall i \in \mathcal{N}, k \in \mathcal{K}^{MEC} \tag{34}$$

$$e_{ik} + c_{ik} = 1 \quad \forall i \in \mathcal{N}, k \in \mathcal{K}^{MEC} \tag{35}$$

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^{MS}_{ik} appearing in Eq. (32) is not counted. Similarly, the amount of CPU consumed by BBU RFBs is computed as:

$$\begin{aligned} C^{BBU}_i &= \textstyle\sum_{k \in \mathcal{K}^{BBU}} \Big[C^{BS}_{ik} d_{ik} + \\ &+ C^{BD}_{ik} \left(\textstyle\sum_{p \in \mathcal{N}} b_{kpi} \sum_{j \in \mathcal{U}} t_{pj} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{36}$$

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

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

$$\sum_{p \in \mathcal{N}} b_{kpi} + f_{ik} \geq 1 \quad \forall i \in \mathcal{N}, k \mathcal{K}^{BBU} \tag{38}$$

[4] We report in Appendix A the extension of the formulation to take into account directed acyclic graphs between set of RFBs of the same type.

$$f_{ik} + d_{ik} = 1 \quad \forall i \in \mathcal{N}, k\mathcal{K}^{BBU} \tag{39}$$

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:

$$C_i^{MEC} + C_i^{BBU} \leq C_i^{CHW} y_i \quad \forall i \in \mathcal{N} \tag{40}$$

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

$$\begin{aligned} M_i^{MEC} &= \textstyle\sum_{k \in \mathcal{K}^{MEC}} \Big[M_{ik}^{MS} c_{ik} + \\ &+ M_{ik}^{MD} \left(\textstyle\sum_{p \in \mathcal{N}} m_{kpi} \sum_{j \in \mathcal{U}} u_{pj} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{41}$$

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

$$\begin{aligned} M_i^{BBU} &= \textstyle\sum_{k \in \mathcal{K}^{BBU}} \Big[M_{ik}^{BS} d_{ik} + \\ &+ M_{ik}^{BD} \left(\textstyle\sum_{p \in \mathcal{N}} b_{kpi} \sum_{j \in \mathcal{U}} u_{pj} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{42}$$

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

$$M_i^{MEC} + M_i^{BBU} \leq M_i^{CHW} y_i \quad i \in \mathcal{N} \tag{43}$$

Finally, Tab. I reports the main notation introduced so far in the optimization problem.

B. Linearization of the Non-Linear Constraints

One of the issues emerging so far is that the constraints (21), (23), (29), (32), (36), (41), (42) are not linear. Non-linear constraints introduce an additional level of complexity. **As a result, it may be challenging to optimally solve the problem** even for small instances. In order to overcome this issue, we detail here the procedure to linearize the non-linear constraints (see [53]). In particular, we can distinguish two kinds of non-linearity: one originated from the product of two binary variables and one associated to the product between a binary and a continuous variable. We focus first on the constraints containing the product between binary variables.

In order to perform a linearization on constraint (21), we define the following binary variables:

$$\xi_{ikj} = r_{ki} u_{ij} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{44}$$

Therefore, the constraint (21) can be re-written as follows:

$$\delta_{ik}^{RRH} = \sum_{j \in \mathcal{U}} \delta_{ikj}^{RRH} \xi_{ikj} \quad \forall i \in \mathcal{N}, k \in \mathcal{K}^{RRH} \tag{45}$$

with the additional following constraints:

$$\xi_{ikj} \leq r_{ki} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{46}$$

$$\xi_{ikj} \leq u_{ij} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{47}$$

$$\xi_{ikj} \geq r_{ki} + u_{ij} - 1 \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{48}$$

Constraints (46) and (47) ensure that ξ_{ikj} is equal to 0 if r_{ik} and u_{ij} are equal to 0. The purpose of the constraint (48) is to guarantee that ξ_{ikj} is equal to 1 if r_{ki} and u_{ij} are equal to 1. Similarly, we can introduce the binary variables:

$$\gamma_{kwip} = r_{ki} b_{wip} \quad i, p \in \mathcal{N}, w \in \mathcal{K}^{BBU}, k \in \mathcal{K}^{RRH} \tag{49}$$

and then we can re-write constraint (29) as follows:

$$\sum_{p \in \mathcal{N}} \gamma_{kwip} \leq O_{kw} \quad \forall i \in \mathcal{N}, k \in \mathcal{K}^{RRH}, w \in \mathcal{K}^{BBU} \tag{50}$$

with the additional constraints:

$$\gamma_{kwip} \leq r_{ki} \quad \forall k \in \mathcal{K}^{RRH}, w \in \mathcal{K}^{BBU}, i, p \in \mathcal{N} \tag{51}$$

$$\gamma_{kwip} \leq b_{wip} \quad \forall k \in \mathcal{K}^{RRH}, w \in \mathcal{K}^{BBU}, i, p \in \mathcal{N} \tag{52}$$

$$\gamma_{kwip} \geq r_{ki} + b_{wip} - 1 \quad \forall k \in \mathcal{K}^{RRH}, w \in \mathcal{K}^{BBU}, i, p \in \mathcal{N} \tag{53}$$

The same strategy can be used also to linearize constraints (41) and (42), by introducing the following two sets of binary variables:

$$w_{kpij} = m_{kpi} u_{pj} \quad i, p \in \mathcal{N}, k \in \mathcal{K}^{MEC}, j \in \mathcal{U} \tag{54}$$

$$\alpha_{kpij} = b_{kpi} u_{pj} \quad i, p \in \mathcal{N}, k \in \mathcal{K}^{BBU}, j \in \mathcal{U} \tag{55}$$

that allow to replace constraints (41) and (42) with the following equivalent equations:

$$\begin{aligned} M_i^{MEC} &= \textstyle\sum_{k\mathcal{K}^{MEC}} \Big[M_{ik}^{MS} c_{ik} + \\ &+ M_{ik}^{MD} \left(\textstyle\sum_{p \in \mathcal{N}} \sum_{j \in \mathcal{U}} w_{kpij} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{56}$$

and

$$\begin{aligned} M_i^{BBU} &= \textstyle\sum_{k\mathcal{K}^{BBU}} \Big[M_{ik}^{BS} d_{ik} + \\ &+ M_{ik}^{BD} \left(\textstyle\sum_{p \in \mathcal{N}} \sum_{j \in \mathcal{U}} \alpha_{kpij} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{57}$$

where the binary variables w_{kpij} have to satisfy:

$$w_{kpij} \leq m_{kpi} \quad \forall i, p \in \mathcal{N}, k \in \mathcal{K}^{MEC}, j \in \mathcal{U} \tag{58}$$

$$w_{kpij} \leq u_{pj} \quad \forall i, p \in \mathcal{N}, k \in \mathcal{K}^{MEC}, j \in \mathcal{U} \tag{59}$$

$$w_{kpij} \geq m_{kpi} + u_{pj} - 1 \quad \forall i, p \in \mathcal{N}, k \in \mathcal{K}^{MEC}, j \in \mathcal{U} \tag{60}$$

and the binary variables α_{kpij} are subject to:

$$\alpha_{kpij} \leq b_{kpi} \quad \forall i, p \in \mathcal{N}, k \in \mathcal{K}^{BBU}, j \in \mathcal{U} \tag{61}$$

$$\alpha_{kpij} \leq u_{ij} \quad \forall i, p \in \mathcal{N}, k \in \mathcal{K}^{BBU}, j \in \mathcal{U} \tag{62}$$

$$\alpha_{kpij} \geq b_{kpi} + u_{ij} - 1 \quad \forall i, p \in \mathcal{N}, k \in \mathcal{K}^{BBU}, j \in \mathcal{U} \tag{63}$$

As regards the remaining constraints, we have to perform the linearization of products between binary and continuous

TABLE I
MAIN NOTATION

	Symbol	Definition	Type - Unit	Appears in Eq.
Sets	$\mathcal{N}$	Set of nodes	-	-
	$\mathcal{U}$	Set of users	-	-
	$\mathcal{K}^{RRH}/\mathcal{K}^{BBU}/\mathcal{K}^{MEC}$	Set of RRH/BBU/MEC RFBs types	-	-
	RRH_k	Set of 5G nodes where an RRH of type k can be placed	-	-
Input Parameters	COV_{ijk}	1 if user j is covered by one RRH RFB of type k installed on node i, 0 otherwise	Boolean	(17)
	N_k^{RRH}	Total number of available RRH RFBs of type k	[units]	(18)
	U_k^{max}	Maximum number of connected users to RRH RFB of type k	[units]	(20)
	δ_{ikj}^{RRH}	Downlink capacity when user j is served by node i with RRH RFB of type k installed on it	[Mbps]	(21),(23)
	R_k^{max}	Maximum capacity that can be handled by an RRH RFB of type k	[Mbps]	(22)
	N_k^{BBU}	Total number of available BBU RFBs of type k	[units]	(26)
	N_k^{MEC}	Total number of available MEC RFBs of type k	[units]	(28)
	O_{kw}	1 if an RRH RFB of type k and a BBU RFB of type w are compatible with each other, 0 otherwise	Boolean	(29)
	δ_k^{MEC}	Total capacity provided by a MEC RFB of type k	[Mbps]	(30)
	δ_k^{BBU}	Total DHW capacity consumed by a BBU RFB of type k	[Mbps]	(31)
	B_i^{DHW}	Amount of installed DHW capacity on node i	[Mbps]	(31)
	C_{ik}^{MS}/C_{ik}^{MD}	Static/Dynamic term for computing the CHW CPU capacity consumed by MEC RFB of type k on node i	[units]/[units/Mbps]	(32)
	C_{ik}^{BS}/C_{ik}^{BD}	Static/Dynamic term for computing the CHW CPU capacity consumed by BBU RFB of type k on node i	[units]/[units/Mbps]	(36)
	$\mathcal{M}$	Very large constant	[units]	(24),(33),(37)
	C_i^{CHW}	Amount of installed CHW CPU resources in node i	[units]	(40)
	M_{ik}^{MS}/M_{ik}^{MD}	Static/Dynamic term for computing the CHW memory consumed by MEC RFB of type k on node i	[units]/[units]	(41)
	M_{ik}^{BS}/M_{ik}^{BD}	Static/Dynamic term for computing the CHW memory consumed by BBU RFB of type k on node i	[units]/[units]	(42)
	M_i^{CHW}	Amount of installed CHW memory resources in node i	[units]	(43)
Variables	u_{ij}	1 if the user j is served by node i, 0 otherwise	Binary	(16),(17),(20),(21),(24),(41),(42)
	r_{ki}	1 if the RRH RFB of type k is installed on node i, 0 otherwise	Binary	(17),(18),(19),(20), (21),(23),(25),(27),(29)
	δ_{ik}^{RRH}	Total capacity to users provided by one RRH RFB of type k installed at node i	[Mbps]	(21),(22),(31)
	t_{ij}	Amount of traffic served to user j by node i	[Mbps]	(23),(24),(30),(32),(36)
	b_{kip}	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	Binary	(25),(26),(29),(31),(36),(37),(38),(42)
	m_{kip}	1 if one MEC RFB of type $k \in \mathcal{K}^{MEC}$ placed at node p is used to serve the RRH RFB at node i, 0 otherwise	Binary	(27),(28),(30),(32),(33),(34),(41)
	y_i	1 if node i is powered on, 0 otherwise	Binary	(31),(40),(43)
	C_i^{MEC}	Amount of CPU capacity consumed on the CHW part of node i by the MEC RFBs installed on it	[units]	(32),(40)
	c_{ik}	1 if at least one MEC RFB of type k is assigned to node i, 0 otherwise	Binary	(32),(33),(35),(41)
	e_{ik}	1 when no MEC RFB of type k is assigned to node i, 0 otherwise	Binary	(34),(35)
	C_i^{BBU}	Amount of CPU capacity consumed on the CHW part of node i by the BBU RFBs installed on it	[units]	(36),(40)
	d_{ik}	1 if at least one BBU RFB of type k is assigned to node i, 0 otherwise	Binary	(36),(37),(39),(42)
	f_{ik}	1 when no BBU RFB of type k is assigned to node i, 0 otherwise	Binary	(38),(39)
	M_i^{MEC}	Amount of memory consumed on the CHW part of node i by the MEC RFBs installed on it	[units]	(41),(43)
	M_i^{BBU}	Amount of memory consumed on the CHW part of node i by the BBU RFBs installed on it	[units]	(42),(43)

variables. To this end, focusing on constraint (23), we define the following continuous variables:

$$\theta_{ikj} = t_{ij} r_{ki} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{64}$$

By means of these new variables we can write constraint (23) in the following equivalent form:

$$\theta_{ikj} \leq \delta_{ikj}^{RRH} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{65}$$

and we add the constraints:

$$\theta_{ikj} \leq \mathcal{M} r_{ik} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{66}$$

$$\theta_{ikj} \leq t_{ij} \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{67}$$

$$\theta_{ikj} \geq t_{ij} - \mathcal{M}(1 - r_{ki}) \quad \forall i \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{RRH} \tag{68}$$

Constraints (66), (67) and (68) guarantee that θ_{ikj} is equal to 0 if r_{ik} is equal to 0 and that θ_{ikj} is equal to t_{ij} if r_{ik} is equal to 1. Analogously, in order to linearize constraint (32) the following new continuous variables are introduced:

$$\phi_{kpij} = m_{kpi} t_{ij} \quad \forall i, p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{MEC} \tag{69}$$

The linearized constraint then follows:

$$\begin{aligned} C_i^{MEC} \;=\; & \textstyle\sum_{k \in \mathcal{K}^{MEC}} \Big[C_{ik}^{MS} c_{ik} + \\ & + \; C_{ik}^{MD} \left(\textstyle\sum_{p \in \mathcal{N}} \sum_{j \in \mathcal{U}} \phi_{kpij} \right) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{70}$$

with the additional constraints:

$$\phi_{kpij} \leq \mathcal{M} m_{kpi} \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{MEC} \tag{71}$$

$$\phi_{kpij} \leq t_{pj} \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{MEC} \tag{72}$$

$$\phi_{kpij} \geq t_{pj} + \mathcal{M}(1 - m_{kpi}) \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{MEC} \tag{73}$$

For the linearization of the constraint (36) we use the same strategy. Then we can define the new following continuous variables:

$$v_{kpij} = b_{kpi} t_{pj} \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{BBU} \tag{74}$$

As a result, constraint (36) can be replaced with the following one:

$$\begin{aligned} C_i^{BBU} &= \textstyle\sum_{k \in \mathcal{K}^{BBU}} \Big[C_{ik}^{BS} d_{ik} + \\ &+ C_{ik}^{BD} \Big(\textstyle\sum_{p \in \mathcal{N}} \sum_{j \in \mathcal{U}} v_{kpij} \Big) \Big] \quad \forall i \in \mathcal{N} \end{aligned} \tag{75}$$

where the variables v_{kpij} are subject to:

$$v_{kpij} \leq \mathcal{M} b_{kpi} \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{BBU} \tag{76}$$

$$v_{kpij} \leq t_{pj} \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{BBU} \tag{77}$$

$$v_{kpij} \geq t_{pj} + \mathcal{M}(1 - b_{kpi}) \quad \forall i,p \in \mathcal{N}, j \in \mathcal{U}, k \in \mathcal{K}^{BBU} \tag{78}$$

C. Overall Formulation

Given the previous definitions of input parameters, variables and linear constraints we pursue as KPIs the maximization of user throughput or the minimization of the number of nodes powered on. We then provide more details about each formulation. Finally, we discuss the classification of the presented formulations.

1) Maximization of user throughput: This KPI aims at maximizing the user performance. More formally, the MAXIMUM USER THROUGHPUT (MAX-UT) problem is defined as:

$$\max \sum_{i,j} t_{ij} \tag{79}$$

subject to:

$$\begin{array}{ll} \text{RRH RFBs Constraints} & (16)-(20),(22),(24) \\ \text{BBU and MEC RFBs Constraints} & (25)-(28),(30) \\ \text{5G Nodes Constraints} & (31),(33)-(35) \\ & (37)-(40),(43) \\ \text{Linearization Constraints} & (45)-(48),(50)-(53), \\ & (56-63),(65-68), \\ & (70)-(73),(75)-(78). \end{array} \tag{80}$$

with control variables: u_{ij}, t_{ij}, r_{ki}, b_{kpi}, m_{kpi}.

2) Minimization of the number of powered on 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 powered on. The following optimization problem is defined:

$$\min \sum_{i} y_i \tag{81}$$

subject to:

$$\begin{array}{ll} \text{RRH RFBs Constraints} & (16)-(20),(22),(24) \\ \text{BBU and MEC RFBs Constraints} & (25)-(28),(30) \\ \text{5G Nodes Constraints} & (31),(33)-(35) \\ & (37)-(40),(43) \\ \text{Linearization Constraints} & (45)-(48),(50)-(53), \\ & (56-63),(65-68), \\ & (70)-(73),(75)-(78). \end{array} \tag{82}$$

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 to set t_{ij} equal to zero for all the users. To avoid this issue, we rely on the ϵ-constrained method of [54] 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 nodes powered on by adding the following constraint to the problem:

$$\sum_{i} y_i \leq N_{used}^{max} \tag{83}$$

where N_{used}^{max} is the maximum number of nodes powered on, 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 set of constraints in (82) and (83) while maximizing the users throughput. We refer to this problem as MINIMUM NUMBER OF NODES (MIN-N).

3) Classification: Both the MIN-UT and MIN-N formulations include the sub-problem of user to RRH RFB association and the RRH RFB link capacity allocation. This sub-problem belongs to the class of Generalized Assignment Problems (GAPs) [55]. However, the MIN-UT and MIN-N formulations include also other constraints, e.g., the BBU and MEC RFBs placement, thus solving a more complex (and challenging) problem.

VII. 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, i.e., when the largest amount of users is requesting a 5G service. Fig. 4 reports the positions of the cells and an example of users placement. 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

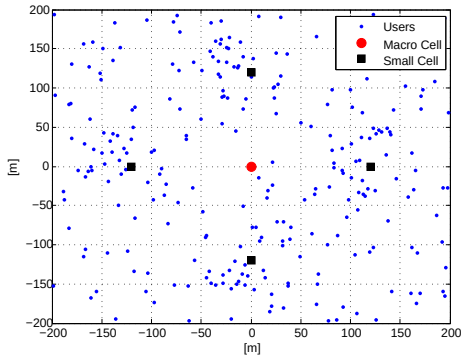


Fig. 4. Reference scenario with macro cell, small cells, and a realization of the users positions.

TABLE II
SUMMARY OF THE SYSTEM PARAMETERS

	Symbol	Value	[Source] / Appear in Eq.
MIMO	N_{OFDM}	7	[51] / Used to compute T_d in Eq. (4)
	τ	3	[51] / Eq. (4)
	T_u	66.7 μs	[51] / Eq. (4)
	T_c	500 μs	[51] / Used to compute T_d in Eq. (4)
	T_s	71.42 μs	[51] / Used to compute T_d in Eq. (4), Appears in Eq. (5)
	T_{slot}	500 μs	[51] / Eq. (5)
	B	20 MHz	[51] / Eq. (5)
	σ	1	[51] / Eq. (5)
	$\frac{(T_{slot}-T_{pilot})}{T_{slot}}$	3/7	[51] / Eq. (5)
	ν	3.8	[51] / Eq. (7)
	w_i^{shad}	8 dB	[51] / Used to compute the z_{ij} terms of Eq. (7)
BBU	S_R	30.72 MHz	[52] / Eq. (11)
	N_B	15	[52] / Eq. (11)
	O_{CW}	16/15	[52] / Eq. (11)
	O_{LC}	10/8	[52] / Eq. (11)

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 the small cell one. Therefore, the resource requirements of the associated RFBs may be different, resulting in two different RFB types.

Tab. II reports the settings of the MIMO and BBU parameters, which rely on the works [51], [52]. In addition, the setting of the RRH RFBs and BBU RFBs parameters is reported in Tab. III, 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

TABLE III
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 IV
5G NODES PARAMETERS

	Parameter	Value		
		Small Cell	Macro Cell	EPC
Capacity	B_i^{DHW}	122.91 [Gbps]	787.91 [Gbps]	727.99 [Gbps]
	C_i^{CHW}	2 [units]	4 [units]	4 [units]
	M_i^{CHW}			
Util.	C_{ik}^{BS},C_{ik}^{MS}	0.5 [units]	0.5 [units]	0.5 [units]
MEC/BBU	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]
	M_{ik}^{BS},M_{ik}^{MS}	0.5 [units]	0.5 [units]	0.5 [units]
	M_{ik}^{BD},M_{ik}^{MD}	0.0116 [units]	0.0019 [units]	0.0019 [units]

with the same type k^*. Focusing then on the BBU RFBs, the BBU parameters of Tab. II 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. III). 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 defined, 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. IV 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

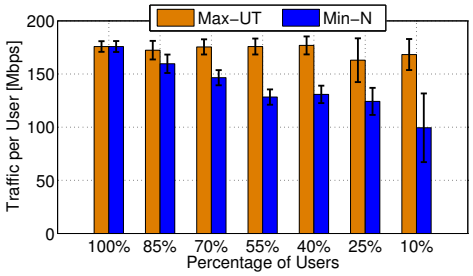


Fig. 5. Traffic per user considering the MAX-UT and MIN-N strategies vs. the percentage of users.

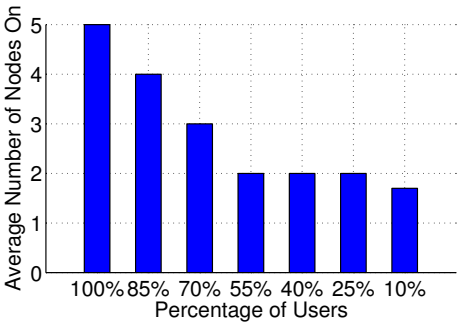


Fig. 6. Average number of nodes on (MIN-N strategy) vs. the percentage of users.

future work.[5] 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).

VIII. 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. Focusing on the scenarios, we initially consider 10 different runs for generating the users' positions. For each run, we then vary the percentage of active users between 100% and 10%. The 100% value corresponds to the peak condition described in Sec. VII. However, the number of users requesting the 5G service normally follows a day-night trend, so it makes sense to evaluate the impact of reducing the percentage of active users to lower values. In particular, we randomly select subsets of users, each of them matching the desired value of percentage. We then solve the MAX-UT and the MIN-N problems over for each scenario run and each value of percentage of users.

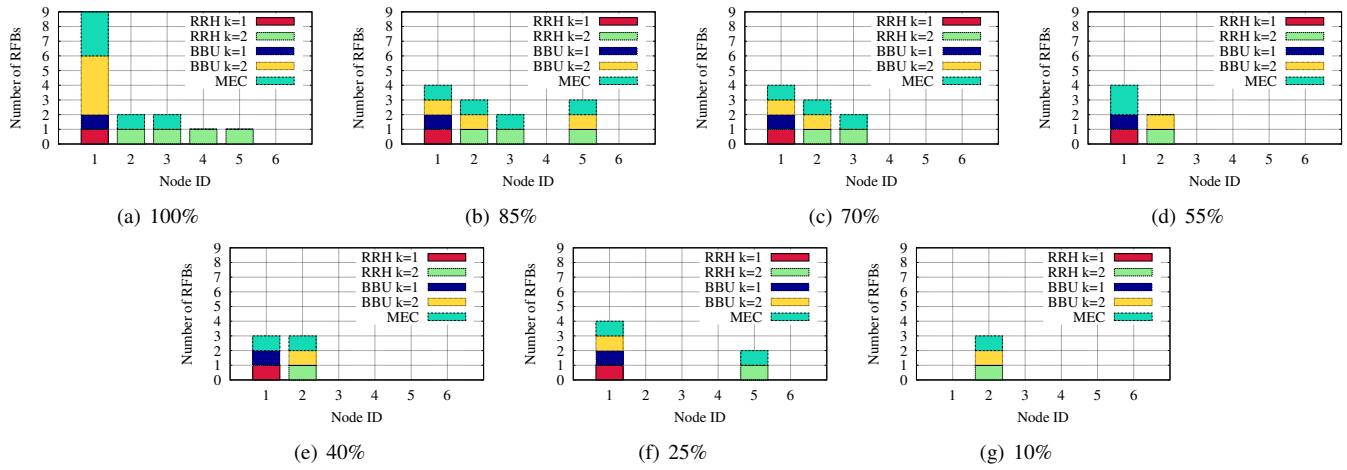
We initially consider the impact on the downlink traffic to users. Fig. 5 reports the traffic assigned on average to each user vs. the variation of the percentage of users for the different strategies. Bars report average values, while the error bars highlight the confidence intervals, which are computed assuming 95% of confidence level. Focusing on the MAX-UT strategy, the average traffic tends to be pretty constant when the percentage of users is decreased. This is due to the fact that, with this strategy, all the 5G nodes are powered on, and therefore they can be used to host RFBs. In this way, the traffic to users tends to be very large, i.e., more than 150 [Mbps] on average. Eventually, a slight decrease in the served traffic is experienced for very low percentages, i.e., 25% and 10%. In particular, when the percentage of users is low, it may happen that the selected subset includes users located at the cells edges, where it is more challenging to properly serve them, due to the bad channel conditions.

Focusing then on the MIN-N strategy, when the percentage of users is equal to 100%, the average traffic is close to the one achieved by the MAX-UT solution. In this case, all the 5G nodes have to be powered on, in order to ensure the user association constraint of Eq. (16), as well as the maximum number of users connected to each RRH RFB of Eq. (20). Then, as soon as the percentage of users is decreased, the average traffic is decreased too, due to the fact that different nodes have been powered off. Nevertheless, thanks to the implemented methodology for solving MIN-N, we can note that, in any case, the average traffic per user is still high, i.e., larger than 100 [Mbps] on average.

In the following, we investigate the effectiveness of the MIN-N strategy in activating a subset of 5G nodes in order to meet the traffic. Fig. 6 reports the results, which are again achieved as an average over the different scenarios for each value of percentage of users. As expected, when the number of users is high, most of 5G nodes have to be powered on. Then, as soon as the percentage of users is lower than 70%, it is possible to keep powered on at most two nodes. At last, when the percentage of the users is low, i.e., equal to 10%, the number of 5G nodes powered on can be even equal to 1 in some cases.

In the next part, we investigate which nodes are actually powered on by MIN-N, and how the RFBs are placed by this strategy. We omit the results obtained by running MAX-UT for two reasons: i) in most cases, they are equivalent to the one of MIN-N when the percentage of users is equal to 100%, ii) the results of MAX-UT do not consistently vary with the change of the percentage of users. Fig. 7 reports the placement of the RFBs for a single scenario, obtained by running MIN-N, and different values of percentages of users. The y-axis reports the cumulative number of RFBs installed in each node, while each bar reports the detail of the RFBs types that are actually installed. Nodes without RFBs installed on them are kept powered off. When the percentage of users is equal to 100%, the Type 1 RRH RFB is installed in the macro cell node,

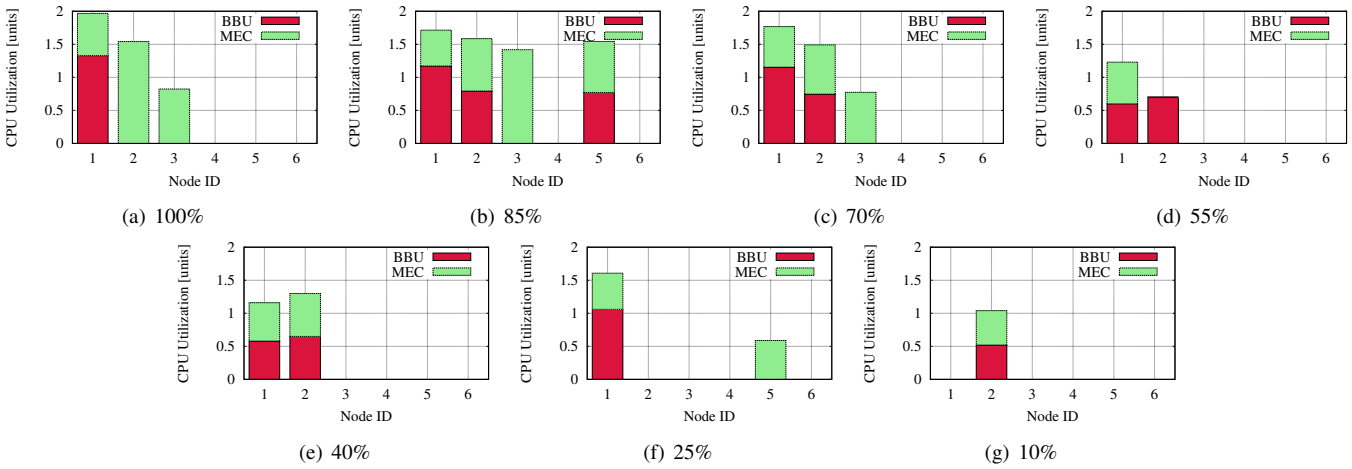
[5]Intuitively, C_i^{CHW} may represent the number of installed CPU cores, while M_i^{CHW} may denote the amount of RAM used.



(a) 100% (b) 85% (c) 70% (d) 55%

(e) 40% (f) 25% (g) 10%

Fig. 7. RFB placement vs. the percentage of users (MIN-N strategy).



(a) 100% (b) 85% (c) 70% (d) 55%

(e) 40% (f) 25% (g) 10%

Fig. 8. Used CHW Capacity in terms of CPU vs. the percentage of users (MIN-N strategy).

while the Type 2 RRH RFBs are installed in all small cell nodes (as expected). Focusing then on the BBU RFBs, they are pooled on the macro cell node. Finally, the MEC RFBs are both installed in the macro cell nodes and two small cell nodes. Then, as soon as the percentage of users is decreased, the number of deployed RFBs is reduced too, thus making possible to realize the required levels of flexibility and dynamicity. In addition, we can note that the BBU and RRH RFBs are moved from node 2 to node 5 when the percentage of users passes from 40% to 25%. This is due to the fact, that, in order to maximize the user traffic, while keeping powered on a subset of nodes, the best option is to deactivate node 2 and to bring the RFB resources on node 5. At last, when the percentage of users is equal to 10%, only one node carrying Type 2 RRH and BBU RFBs is able to satisfy the user connectivity and to maximize the traffic.

Up to this point, a natural question is then: What is the impact of the proposed strategies on the CHW resources? To investigate this issue, we report in Fig. 8 the amount of used CHW capacity in terms of CPU obtained by running the MIN-N strategy over a scenario.[6] The subfigures report the results by varying the percentage of users. In general, the amount of consumed CPU is approximately equal to 50% of the maximum one (which we recall is equal to 2 [units] for each small cell and 4 [units] for the macro cell). From the figures, we can note that the amount of consumed CPU is not constant, but heavily depends on the number of users served, as well as the number of powered on nodes. In particular, we recall that the amount of consumed CPU depends on a static term, which has to be counted if at least one BBU/MEC RFB is installed in the node, plus a dynamic term, that instead depends on the amount of traffic served to users. Finally, we can observe that there are some nodes installing only BBU RFBs or only MEC RFBs.

In the last part of our work, we have measured the computation time of the MIN-N over the different scenarios. Fig. 9 reports minimum, average and maximum computation times for each percentage of users. As expected, when the percentage of users is decreased, the computation time is also decreased, due to the fact that the problem is smaller in

[6]Similarly to the previous cases, the MAX-UT results are similar to the ones of the MIN-N strategy when the percentage of users is equal to 100%.

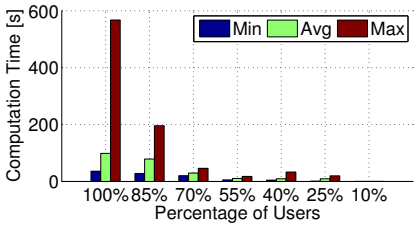


Fig. 9. Computation time vs. the percentage of users (MIN-N strategy).

terms of number of variables and constraints. Moreover, we can note that the decrease is more than linear (especially for the maximum values). Overall, these numbers suggest that the optimal solution approach can be feasible when the number of users and the number of nodes is in the same order of magnitude than the ones of the scenarios considered in this work. Eventually, when the size of the scenario is further increased, a sub-optimal approach should be preferred, in order to reduce the computation times. We leave the investigation of this aspect as future work.

IX. Conclusions and Future Works

We have targeted the management of the RFBs in a superfluid 5G network, as an outcome of the SUPERFLUIDITY project. We have considered different KPIs, including the maximization of the users traffic, and the minimization of the number of used nodes. After detailing the RFBs features, as well as the characteristics of the physical nodes hosting them, we have proposed a set of models to define the RFBs requirements in terms of CHW (CPU and memory) and DHW resources, as well as a model for the HW of the 5G node. We have then formulated the problem of managing a set of RFBs in a 5G network under the aforementioned KPIs. After showing that different constraints of the problem are not linear, we have then detailed how these constraints are linearized, in order to reduce the problem complexity. Our results, obtained by solving the linearized problem over different representative scenarios, confirms that the proposed RFBs-based approach allows a high level of flexibility and dynamicity, coupled with an extremely good performance to users. In particular, we have shown that the RFBs can be efficiently moved across the set of nodes, being able to realize the required service to users. As a result, the user throughput is larger than 150 [Mbps] when the maximization of the user experience is pursued, and larger than 100 [Mbps] on average when the goal is the reduction of the number of 5G nodes powered on.

As next steps, we plan to face different issues, including: i) the investigation of more detailed channel models, ii) the consideration of smaller RFBs (in terms of deployed functionalities), as well as the introduction of more complex relationships between the RFBs, and iii) the design of new algorithms to solve the problem even for large instances composed of thousands of users and hundreds of RFBs.

Acknowledgments

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

References

[1] L. Chiaraviglio, L. Amorosi, S. Cartolano, N. Blefari-Melazzi, P. Dell'Olmo, M. Shojafar, and S. Salsano, "Optimal superfluid management of 5g networks," in *Network Softwarization (NetSoft), 2017 IEEE Conference on*, pp. 1–9, IEEE, 2017.

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

[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] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5g wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, 2014.

[11] M. Hasan and E. Hossain, "Distributed resource allocation in 5g cellular networks," *Towards 5G: Applications, Requirements and Candidate Technologies*, pp. 129–161, 2014.

[12] X. Ge, H. Cheng, M. Guizani, and T. Han, "5g wireless backhaul networks: challenges and research advances," *IEEE Network*, vol. 28, no. 6, pp. 6–11, 2014.

[13] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!," *IEEE access*, vol. 1, pp. 335–349, 2013.

[14] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[15] S. Han, L. Dai, Q. Sun, Z. Xu, *et al.*, "Full duplex networking: mission impossible?," *arXiv preprint arXiv:1410.5326*, 2014.

[16] A. Agrawal, "Heterogeneous networks. a new paradigm for increasing cellular capacity," *Qualcomm, Jan*, vol. 29, 2009.

[17] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE pervasive Computing*, vol. 8, no. 4, 2009.

[18] F. Liu, P. Shu, H. Jin, L. Ding, J. Yu, D. Niu, and B. Li, "Gearing resource-poor mobile devices with powerful clouds: architectures, challenges, and applications," *IEEE Wireless communications*, vol. 20, no. 3, pp. 14–22, 2013.

[19] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, ACM, 2012.

[20] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile Edge Computing - A key technology towards 5G," *ETSI White Paper No. 11*, 2015.

[21] ETSI MEC ISG, "Mobile Edge Computing (MEC); Framework and Reference Architecture." ETSI GS MEC 003 V1.1.1, 2016.

[22] D. Sabella, A. Vaillant, P. Kuure, U. Rauschenbach, and F. Giust, "Mobile-edge computing architecture: The role of MEC in the Internet of Things," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 84–91, 2016.

[23] Alex Reznik (editor), "Developing Software for Multi-Access Edge Computing," *ETSI White Paper No. 20*, 2017.

[24] Intel and Nokia Siemens Networks, "Increasing Mobile Operators Value Proposition With Edge Computing." Technical Brief, 2013.

[25] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating While Computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 45–55, 2014.

[26] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, 2015.

[27] J. Sánchez, I. G. B. Yahia, N. Crespi, T. Rasheed, and D. Siracusa, "Softwarized 5g networks resiliency with self-healing," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, pp. 229–233, IEEE, 2014.

[28] S. Kukliński, K. T. Dinh, and T. Rasheed, "MCCN: A Softwarized Approach to 5G,"

[29] I. Giannoulakis, E. Kafetzakis, G. Xylouris, G. Gardikis, and A. Kourtis, "On the Applications of Efficient NFV Management Towards 5G Networking," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, IEEE, 2014.

[30] C. Bouras, P. Ntarzanos, and A. Papazois, "Cost Modeling for SDN/NFV Based Mobile 5G Networks," in *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2016 8th International Congress on*, pp. 56–61, IEEE, 2016.

[31] A. A. Gebremariam, T. Bao, D. Siracusa, T. Rasheed, F. Granelli, and L. Goratti, "Dynamic strict fractional frequency reuse for software-defined 5g networks," in *Communications (ICC), 2016 IEEE International Conference on*, pp. 1–6, IEEE, 2016.

[32] S. Mehraghdam, M. Keller, and H. Karl, "Specifying and placing chains of virtual network functions," in *Cloud Networking (CloudNet), 2014 IEEE 3rd International Conference on*, pp. 7–13, IEEE, 2014.

[33] H. Moens and F. De Turck, "VNF-P: A model for efficient placement of virtualized network functions," in *Network and Service Management (CNSM), 2014 10th International Conference on*, pp. 418–423, IEEE, 2014.

[34] B. Addis, D. Belabed, M. Bouet, and S. Secci, "Virtual network functions placement and routing optimization," in *Cloud Networking (CloudNet), 2015 IEEE 4th International Conference on*, pp. 171–177, IEEE, 2015.

[35] A. Gupta, M. F. Habib, P. Chowdhury, M. Tornatore, and B. Mukherjee, "Joint virtual network function placement and routing of traffic in operator networks," *UC Davis, Davis, CA, USA, Tech. Rep*, 2015.

[36] M. Dieye, S. Ahvar, J. Sahoo, E. Ahvar, R. Glitho, H. Elbiaze, and N. Crespi, "Cpvnf: Cost-efficient proactive vnf placement and chaining for value-added services in content delivery networks," *IEEE Transactions on Network and Service Management*, 2018.

[37] D. S. Michalopoulos, M. Doll, V. Sciancalepore, D. Bega, P. Schneider, and P. Rost, "Network slicing via function decomposition and flexible network design," 2017.

[38] V.-G. Nguyen, A. Brunstrom, K.-J. Grinnemo, and J. Taheri, "Sdn/nfv-based mobile packet core network architectures: a survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1567–1602, 2017.

[39] A. Gopalasingham, D. G. Herculea, C. S. Chen, and L. Roullet, "Virtualization of radio access network by virtual machine and docker: Practice and performance analysis," in *Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium on*, pp. 680–685, IEEE, 2017.

[40] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, and S. Gosselin, "Optimal bbu placement for 5g c-ran deployment over wdm aggregation networks," *Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1963–1970, 2016.

[41] S.-H. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7621–7632, 2016.

[42] L. Chiaraviglio, F. DAndreagiovanni, G. Siderotti, N. B. Melazzi, and S. Salsano, "Optimal Design of 5G Superfluid Networks: Problem Formulation and Solutions," in *21st Conference on Innovation in Clouds, Internet and Networks (ICIN) 2018*, 2018.

[43] "Building service testbeds on fire." (Date last accessed Feb 2017).

[44] "Felix project (federated test-beds for large-scale infrastructure experiments)." (Date last accessed Feb 2017).

[45] "Selfnet: A framework for self-organized network management in virtualized and software defined networks." (Date last accessed Feb 2017).

[46] P. Neves, R. Calé, M. R. Costa, C. Parada, B. Parreira, J. Alcaraz-Calero, Q. Wang, J. Nightingale, E. Chirivella-Perez, W. Jiang, *et al.*, "The selfnet approach for autonomic management in an nfv/sdn networking paradigm," *International Journal of Distributed Sensor Networks*, 2016.

[47] "ETSI GS NFV-SWA 001: Network Functions Virtualisation (NFV); Virtual Network Functions Architecture, V 1.1.1," *ETSI, December*, 2014.

[48] G. Bianchi, M. Bonola, A. Capone, and C. Cascone, "Openstate: programming platform-independent stateful openflow applications inside the switch," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 2, pp. 44–51, 2014.

[49] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek, "The click modular router," *ACM Transactions on Computer Systems (TOCS)*, vol. 18, no. 3, pp. 263–297, 2000.

[50] M. Bansal, J. Mehlman, S. Katti, and P. Levis, "Openradio: a programmable wireless dataplane," in *Proceedings of the first workshop on Hot topics in software defined networks*, pp. 109–114, ACM, 2012.

[51] 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.

[52] 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.

[53] W. P. Adams and H. D. Sherali, "Linearization strategies for a class of zero-one mixed integer programming problems," *Operations Research*, vol. 38, no. 2, pp. 217–226, 1990.

[54] 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.

[55] R. M. Nauss, "Solving the generalized assignment problem: An optimizing and heuristic approach," *INFORMS Journal on Computing*, vol. 15, no. 3, pp. 249–266, 2003.

APPENDIX A
INTRODUCING DIRECTED ACYCLIC GRAPHS

The presented formulations are able to manage the RFBs in a very flexible way, by assuming that each RFB of the same type k is able to be run independently by the other RFBs of the same type. However, the operator may be willing to impose that a BBU serves multiple RRHs. As a result, the RFBs of the same type may be constrained to be placed on the same node, e.g., in order to take advantage of the sharing of the physical node hosting them. This is translated in the general case with the creation of directed acyclic graphs rather than chains. To mimic this behavior, we need to introduce a set of additional constraints in our formulation. In particular, let us consider the case in which all the BBU RFBs of the same type k have to be co-located together on the same node. In a similar way, let us assume that the MEC RFBs of the same type k need to be co-located together on the same node (which may be different from the one used to host the BBU RFBs). We impose these two conditions by adding the following set of constraints:

$$\beta_{kp} \geq b_{kip} \quad \forall k \in \mathcal{K}^{BBU}, i, p \in \mathcal{N} \tag{84}$$

$$\sum_i b_{kip} = N_k^{BBU} \beta_{kp} \quad \forall p \in \mathcal{N}, k \in \mathcal{K}^{BBU} \tag{85}$$

$$\mu_{kp} \geq m_{kip} \quad \forall k \in \mathcal{K}^{MEC}, i, p \in \mathcal{N} \tag{86}$$

$$\sum_i m_{kip} = N_k^{BBU} \mu_{kp} \quad \forall p \in \mathcal{N}, k \in \mathcal{K}^{MEC} \tag{87}$$

where β_{kp} and μ_{kp} are binary variables equal to 1 if one BBU (MEC) of type k is located on p (0 otherwise).