Perez, G. O., Ebrahimzadeh, A., Maier, M., Hernandez, J. A., Lopez, D. L., Veiga, M. F. (2020). Decentralized Coordination of Converged Tactile Internet and MEC Services in H-CRAN Fiber Wireless Networks. *Journal of Lightwave Technology*, 38(18), 4935–4947.

DOI: https://doi.org/10.1109/jlt.2020.2998001

Decentralized Coordination of Converged Tactile Internet and MEC services in H-CRAN Fiber Wireless Networks

Gabriel Otero Pérez*, Amin Ebrahimzadeh†, Martin Maier†,
José Alberto Hernández*, David Larrabeiti López*, and Manuel Fernández Veiga‡
*Department of Telematics Engineering, Universidad Carlos III de Madrid, Spain
†Institut National de la Recherche Scientifique (INRS), Canada
‡Universidade de Vigo, Spain

***Abstract*—In order to meet the requirements of services and applications envisioned for post-5G and 6G networks, research efforts are heading towards the convergence of architectures aiming to support the wide variety of new compute-demanding and latency-sensitive applications in the context of Tactile Internet. In this paper, we study the resource allocation and association of users with different delay requirements in a shared-backhaul fiber-wireless (FiWi) enhanced Heterogeneous Cloud Radio Access Network (H-CRAN) with Multi-access Edge Computing (MEC) and offloading. As opposed to traditional resource and association management, we propose a decentralized algorithm based on a full dual decomposition of the optimization problem to operate the network. Results show that this approach outperforms the traditional one in terms of average delay and energy consumption, achieving up to 80% average delay improvement in high-load scenarios.**

***Index Terms*—Backhaul Awareness, Downlinl-Uplink Decoupling, Fiber-Wireless, Heterogeneous Cloud Radio Access Network, Self-Organizing Networks.**

I. Introduction

The upcoming 5G cellular networks are expected to cope with a massive increase in the number of mobile smart devices and new applications. Research efforts have focused on achieving massive bandwidth, extreme densification and offloading, ultra-low round-trip times, and improved spectral efficiency [1]. Also, the uplink has gained importance due to the emerging Internet of Things (IoT).

One of the promising technologies to achieve the abovementioned requirements and reduce CAPEX and OPEX of future netwoks is the Cloud Radio Access Network (C-RAN) architecture. C-RAN moves the processing of cellular radio signals from base stations to the cloud. Accordingly, conventional base stations are split into three elements: Central Unit (CU), Distributed Unit (DU), and Remote Radio Unit (RRU). The latter is in charge of radiating radio signals to the users as well as gathering and digitizing them back to the processing units by using appropriate transport protocols like eCPRI over IP or Ethernet [2]. Additionally, DU and CU share further processing functions that are dependent on the particular deployment [3]. This network architecture obviously reduces the complexity and cost of the deployed base stations. It supports more complex and smarter Coordinated Multi-Point (CoMP) services, better interference management, energy-efficient cooling and virtualization features that enable an agile and fast introduction of new services [4]. However, this scheme poses stringent latency requirements on the digitized data that must be met for the proper functioning of the network.

Recently, a combination of C-RAN and heterogeneous networks (HetNets), known as Heterogeneous Cloud Radio Access Network (H-CRAN) [5], has emerged featuring several types of base stations. Compared to C-RAN, H-CRAN makes use of LTE-A and WiFi to alleviate the burden on the fronthaul links and support offloading through different technologies. In this context, a careful design of the network becomes paramount to ensure that both C-RAN and non-C-RAN/backhaul data can coexist while meeting the requirements for both. Moreover, there exists the desire of going beyond this coexistence and achieving the convergence of technologies that brings us the best of both architectures and services.

Conventional communications between human mobile users is not the only type of traffic that is envisioned in post-5G/6G networks. In this paper, we consider the coexistence of three types, each with a particular set of quality-of-service (QoS) requirements: (1) *Human-to-Human* (H2H) communications with moderate delay requirements, (2) *Human-to-Machine* (H2M) communications, in the so-called Tactile Internet, posing stringent delay requirements for *immersive* services and applications such as telepresence, remote control of haptic machines, and remote surgery, and (3) *Multi-access Edge Computing* (MEC) traffic, providing cloud computing capabilities at the edge of access networks, leveraging the physical proximity of servers and mobile users to achieve a reduced latency and increased reliability.

Unfortunately, 5G deployments expose limitations regarding the integration of new applications [6]. To overcome this, there exists an increasing interest in next-generation 6G systems focused on the convergence of technologies [7]. The major research branches of 6G consider not only delivering another 1000x increase in data rates, but also diving into self-sustaining networks and dynamic resource utilization. 6G will also put an end to smartphone-centric networks, introducing new system paradigms (e.g., human-in-the-loop communications, human-centric services). In this paper, we aim at solving the user association and allocation of communication and computation resources to different types of users over backhaul-shared fiber-wireless (FiWi) enhanced H-CRAN networks supporting three transmission technologies: LTE-A, WiFi, and C-RAN RRUs. Convergence in 6G appears to be achievable only at the expense of an increased complexity. To cope with this, decentralization represents a promising means to keep the scalability via self-organization.

Moreover, the objective of this paper is to develop a decentralized algorithm with very low information pass requirements that is able to accomodate users (humans, machines and computation tasks) to base stations of different technologies (WiFi, LTE-A, C-RAN) in an optimal way so as to fulfill their wide range of latency requirements and maximize the global network utility.

The remainder of the paper is structured as follows. Section II briefly reviews the related work. In Section III, we present the network model based on a two-tier heterogeneous H-CRAN network supported by an Ethernet Passive Optical Network (EPON) backhaul, trying to answer several interesting migration questions such as how to integrate C-RAN in decentralized HetNets or the evolution of future C-RAN. Additionally, we introduce a signal model and a traditional association scheme based on stochastic geometry that will serve as a baseline to compare with our derived solution. Section IV presents the optimization problem and a scalable decentralized solution is shown in Section IV-B based on a full dual decomposition of the optimization problem. We further validate the simulator in Section V-A. We show the utility of our proposed approach in the use case of the emerging *Tactile Internet* in Section V-B. Finally, Section VI summarizes the conclusions of the paper.

II. Related Work

Optimization of **cellular networks** through association biasing [40], balancing between delay and throughput [41], [42], or cell breathing [43], have been major research topics in HetNets as a solution to maximize throughput and improve latency. In [44], the authors present a holistic interference optimization framework for LTE-A HetNets.

There already exist examples proposing the design mobility-aware association strategies to overcome the limitations of the conventional received power-based association schemes as in [29]. Even though an interesting approach is followed to reduce frequent handovers, the optimization algorithm does not guarantee delay requirements or differentiate between types of traffic. In [30], the user association for ultra-low latency in LTE-A is studied from a Bayesian cell selection and usser association perspective. The authors of [45] propose a cooperative strategy based on centralized resource scheduling and statistical traffic data. In [46], a taxonomy of different association algorithms is presented. This extensive overview reviews the most common options used for heterogeneous networks, massive MIMO, mmWave, and energy harvesting networks.

More recently, the authors of [47] have explored the benefits of opportunistic relaying to achieve load balancing and throughput gains in 5G/6G vehicular wireless networks. In [48], the authors study an energy-efficiency optimization problem in **H-CRAN networks**. Authors of [49] solve an energy and bandwidth consumption minimization problem for an H-CRAN network to decide on the used functional split for each service. However, the C-RAN network is usually not integrated with the rest of the network [5], using its own aggregation network.

The use of EPON to support cloud and edge computing capabilities has caught attention in the past [50], [51], studying the suitable EPON features needed to provide these services. The cloud vs **MEC** collaborative offloading is studied in [36] using a game-theoretic framework. Recently, Castellano et al. [52] studied the optimal assignment of computing resources for heterogeneous edge applications. Finally, the use of priority queuing to ensure that the **C-RAN** requirements are met, is a common practice [12].

Other studies, rely on guaranteeing extreme delay percentiles [2], [10]. To the best of our knowledge, there is a lack of solutions that analyze the multi-delay-constrained joint user association and resource allocation in H-CRAN networks while solving important network convergence and design aspects to achieve a true operational integration.

III. System Model

A. Network Architecture

Figure 1 depicts the generic architecture of the considered FiWi enhanced H-CRAN network. The fiber backhaul consists of a time or wavelength division multiplexing (TDM/WDM) IEEE P802.3ca 25 Gbps EPON [8] with a typical fiber range of 10 km between the central optical line terminal (OLT) – located at the Central Office (CO)– and remote optical network units (ONUs), which may be extended up to 100 km to account for long-reach PON deployment scenarios. The EPON may comprise multiple stages, each separated by a wavelength broadcasting splitter/combiner or a wavelength multiplexer/demultiplexer.

There are four different subsets of ONUs. An ONU may serve fixed (wired) subscribers, it may connect to a cellular LTE-A base station (BS) or an IEEE 802.11n/ac/s WLAN mesh portal point (MPP), giving rise to a collocated ONU-BS or ONU-MPP, respectively. Finally, ONUs are co-located with C-RAN DUs that aggregate and preprocess fronthaul data coming from the RRUs, creating what we call the ONU-DUs.

Note, in Figure 1, the implemented color code is used to indicate the coverage areas of each technology. We use orange color to fill the coverage areas that can be served using WiFi technology, and blue color to indicate that 5G C-RAN service is available in a given region. Finally, the areas filled with green color represent those that can be served via LTE-A technology. It is worth highlighting that, since the LTE-A coverage areas are larger than those of WiFi and 5G C-RAN, these are surrounded by the former. This means that some users (humans or robots shown in the figure) might have

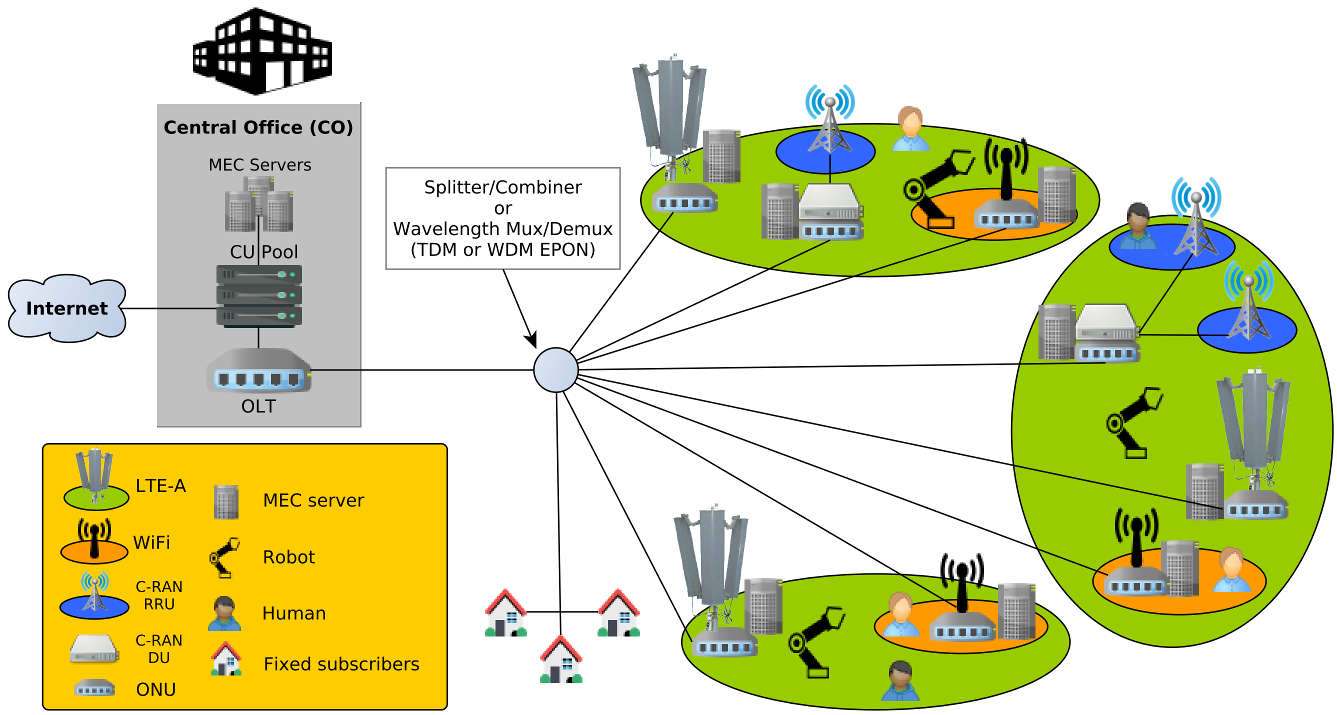


Fig. 1. Generic architecture of the MEC-enabled FiWi enhanced H-CRAN with coexistent C-RAN, H2H, and H2M traffic.

several options regarding the technology they use to access the network.

Depending on her trajectory and position, a mobile user (MU) may communicate through the cellular network and/or WLAN mesh front-end, which consists of ONU-MPPs, intermediate mesh points (MPs), and mesh access points (MAPs). Alternatively, a MU may access the network through the C-RAN deployment. The C-RAN RRUs are in charge of collecting, digitizing, and sending the MU's signals that will be processed at the DU and CU.

In this architecture, DUs and RRUs are deployed close to each other (on the order of 100s of meters) connected using fiber [3], [9] or Ethernet point-to-point links [10]. However, the unprocessed fronthaul traffic poses stringent delay requirements as noted in the eCPRI and IEEE 802.1CM time-sensitive networking for fronthaul protocols [11], [12]. In summary, eCPRI and the IEEE Standard identify the IQ data sent by the RRUs as high priority fronthaul (HPF) class, with a hard budget limit for the end-to-end one-way latency. Assuming that RRUs handle a low layer functional split (e.g., Option 7 Intra-PHY - eCPRI Split [9]) a protocol like eCPRI may be utilized to encapsulate the digital samples and send them to the DU for further processing.

After the processing is performed at the DU, the latency requirements imposed on the EPON are relaxed as we move to higher layer splits like Split Option 2 (PDC/high RLC [13]), and thus enabling the network to transport C-RAN traffic with a latency budget between and ms for Split Option 2 [14]. Further explanations of the functions placed at each processing element can be found in Section III-D. Finally, after traversing the EPON, the traffic reaches the CU pool at the OLT, where the final processing takes place.

From the network coverage and operational point of view, a two-tier multi-cell H-CRAN network is considered with macro-cells providing LTE-A connectivity and small cells featuring low-power and low-cost offloading. These small cells are realized via WiFi MAPs and C-RAN RRUs. Note that only few of them are highlighted in Figure 1 for illustration, using orange, and blue color, respectively, to mark their coverage areas. Through densification, users are virtually able to associate with any radio technology, going beyond the usual cell boundaries imposed by traditional power criteria. The management data (control and signaling) and user data are treated in different ways.

User traffic is transported using the infrastructure of the user's host network. However, taking advantage of the macro-cell coverage, the control/signaling and broadcast data are sent via the LTE-A macro-cells to the users, including the C-RAN management information. Later in this paper, once the optimization framework takes control of the network operation, the LTE-A network can also be used to broadcast the appropriate information that will enable the users to achieve self-organization. Consequently, RRUs are mere hot-spots used to provide high-speed data transfer. This also helps relax the time and delay constraints in the fronthaul network located between the CU/DU and the RRUs and thereby enables a truly cooperative operation between C-RAN and non-C-RAN networks, since the signaling data of both networks are no longer unaware of each other.

In our work, several types of communication are taken into account. In addition to conventional human end-users with their typical triple-play human-to-human (H2H) traffic (i.e., voice, video, and data), we allow for the coexistence of Tactile Internet robots, paired with human operators (HO), as the next evolutionary step of IoT. Telesurgery is a well-known application of such robots, featuring a number of sensors and manipulators with different degrees of freedom (DoF), typically ranging from 3 to . Samples containing the updated positions and orientation signals are transmitted from the HO to the robot, and vice versa. These samples must comply with certain QoS guarantees for an adequate immersive experience, i.e., to extend human capabilities rather than just automate them.

Finally, we equip ONU-BSs/MPPs/DUs with MEC servers – simply called *edge servers* hereafter– collocated at the optical-wireless interface, as shown in Figure 1. Therefore, MUs associated to LTE-A, WiFi, or C-RAN may offload their incoming computational tasks to nearby edge servers or execute their computation tasks. Alternatively, a user may decide to offload a task to an MEC server located at the CO [15] for more powerful computation.

B. Radio signal model

We assume a two-tier HetNet where all BS in tier , , are located according to two independent homogeneous Poisson Point Processes (PPP) of density . Let be the transmission power of a node at tier , where for LTE-A BSs and for WiFi MAPs and RRUs. The transmission power for each UE is . We consider additive white Gaussian noise with constant power . Thus, the received power in the downlink channel measured at a typical user device located at $\mathbb{R}$ and receiving a signal from a BS located at is

(1)

where is the Euclidean norm, is the path loss exponent, and is the small-scale Rayleigh fading. Similarly, the signal power received at the BS in the uplink is given by

(2)

We assume some form of orthogonal transmission (e.g., TDMA or OFDMA) so that no intra-cell interference exists. However, interference from distant transmitters at other cells cannot be neglected, especially when the path-loss exponent is low ( ). The resulting downlink SINR expression assuming a user device at connects to a BS in tier located at position is

$$\mathsf{SINR} \qquad \frac{\qquad}{\qquad}$$

(3)

where is the -th tier. The same assumptions hold for the uplink transmissions.

C. Traditional Association Scheme

In traditional cellular networks, it has been almost a universal practice to associate an MU in both uplink and downlink to the same BS. Downlink/Uplink Decoupling (DUDe) represents an alternative idea that proposes the use of different BSs for uplink/downlink [16]–[18].

Current LTE-A networks enable dual connectivity from the MUs, and DUDe has been shown to outperform coupled association schemes [19], [20] besides having architectural advantages, such as reliability. In DUDe, the association decision is now based on the average received signal in DL/UL separately, taking the expectation over the probability density function of the fading process. We can obtain the received signal powers for uplink and downlink by averaging (1) and (2)

$$\mathbb{E} \qquad \mathbb{E}$$

where stands for the tier. Let be the distance between the device and the serving base station at tier . Thus, under the maximum-received-power policy, the following association rules can be derived: Connect to an LTE-A BS in the downlink if , otherwise connect to a tier 2 BS. Similarly, associate to an LTE-A BS in the uplink if and associate to tier 2 otherwise. However, these decisions are independent from each other and lack a holistic view of the system. Clustering the users according to these rules leads to the so-called Voronoi coverage maps, built under the premise of maximizing the received power and minimizing the transmission distances.

Nevertheless, it is clear that this is not always optimal, as load imbalance takes place due to the disparate transmit powers of base station. Additionally, despite the fact that DUDe still yields substantial performance gains over co-located association, fixed coverage maps may cause that some BSs are serving most of the users, whereas some others are idle. Even though this can be partially solved by increasing the BS densification, other approaches are investigated in more detail in the following sections via the optimization theory.

D. Traffic Generation and Requirements

One characteristic of the service ecosystem envisioned for future mobile networks is heterogeneity. For **H2H communications**, we assume that the packet size of a mobile user is $_{\text{H2H}}$. Also, this data has moderate delay requirements, given by the maximum end-to-end delay: $^{\text{Threshold}}_{\text{H2H}}$. Regarding the **H2M traffic**, we assume that a human operator (HO) in a haptic session of a teleoperation system with degrees of freedom requires an end-to-end delay lower than $^{\text{Threshold}}_{\text{H2M}}$ (usually, sub ms [21]) as well as a packet size of $_{\text{H2M}}$.

With respect to **MEC traffic**, tasks are characterized by an uplink and downlink packet size $_{\text{MEC}}$ and $_{\text{MEC}}$. Also, we assume these tasks have a maximum completion time of $^{\text{Threshold}}_{\text{MEC}}$. In the baseline scenario, each MU uses a scheduler to decide whether to offload a task to an Edge/CO server or execute it locally, with probability $_{\text{Offload}}$. If the decision is to offload, it will decide between Edge/CO with probability $_{\text{Edge}}$ and $_{\text{Edge}}$, respectively (only applies to MEC users). Conversely, the distributed solution will take care of the decision in the proposed approach.

Wired subscribers are assumed to contribute to the EPON's occupancy with average background rates of ${}^{\text{UL}}_{\text{Wired}}$ and ${}^{\text{DL}}_{\text{Wired}}$.

C-RAN RRUs will periodically send traffic to the DU (and to the CU pool, ultimately) so that the signal processing can be performed on the received signals. This traffic has stringent delay requirements, as low as ${}^{\text{RRU-DU (Max)}}_{\text{C-RAN}}$. The size of the packets and bursts highly depends on the eCPRI functional split we choose to build our network [2]. In this paper, we choose Split ${}_{\text{U}}$, which represents the radio signals just after removing the cyclic prefix and performing the Fast Fourier Transform, getting rid of the unused guard band subcarriers, and demapping the resource blocks. Note that all the processing functions we just mentioned are performed at the RRU, that is down to Option 7 Intra-PHY, not included. This split particularly matches [3] the CU/DU distribution explained in Section III. The output traffic of a single RRU can be characterized as

$$\text{Split } \mathbf{I}_{\text{U}} \qquad {}_{\text{sc}} \tag{4}$$

where ${}_{\text{sc}}$ defines the used number of subcarriers, is the symbol period, and , , and stand for the cell utilization, number of bits used to quantize signals and the number of antennas, respectively. Note that the cell utilzation ( ) is directly affected by the number of users connected to the antenna. Since represents the fraction of resource blocks that are under use, the more users in the cell, the more resources are occupied.

Assuming a 4-antenna MIMO system supporting MHz channels with KHz subcarrier spacing, the number of used subcarriers is . To maintain orthogonality, the symbol interval is $\overline{}$. Using bits to quantize the signals and assuming a worst-case utilization of , the resulting rate for one RRU is Mbit/s. In other words, a burst of ${}^{\text{RU-DU}}_{\text{C-RAN}}$ bytes is sent every $\overline{}$.

After the fronthaul traffic reaches the DU, this processing element must perform all the tasks included in Option 7 split, down to split Option 2 (F1 interface). Among other functions, between these splits we can find tasks such as the channel estimation, equalization, demodulation, descrambling, decoding, etc. These will be performed at the DU.

Once that this traffic leaves the DU towards the CU, the capacity requirements are relaxed since overheads are processed and eliminated. At this moment, we are at split Option 2 (F1 interface) functional split [9], as mentioned in Section III-A. The authors of [22] calculate the fronthaul rate for a MHz channel after reaching Option 2 processing split (see Table 1 in [22] and Appendix C of [23]). They assume a number of MHz chunks, up to a total of MHz. Together with state-of-the-art assumptions about the modulation and number of layers, the total radio bandwidth accounts for a fronthaul rate of Mb/s in the downlink and Mb/s in the uplink (see Table A-1 of [14]). We use these estimations for the DU-CU transmission rates through the EPON. The rest of the processing steps will be performed at the CU site.

E. End-To-End Delay

In our delay analysis, we take into consideration different ways of computing the end-to-end delay perceived by a user, depending of what network(s) they make use of.

In **Cellular and WiFi commnunications**, the transmission delays in uplink and downlink are modeled, for each user, as

$$ {}^{\text{UL}}_{\text{Comm}} \qquad \frac{{}^{\text{UL}}_{\text{H2H}}}{{}_{\text{UL}}} \tag{5}$$

$$ {}^{\text{DL}}_{\text{Comm}} \qquad \frac{{}^{\text{DL}}_{\text{H2H}}}{{}_{\text{DL}}} \tag{6}$$

where ${}^{\text{UL}}$ and ${}^{\text{DL}}$ are the packet sizes for uplink and downlink, and ${}_{\text{UL}}$ and ${}_{\text{DL}}$ are the uplink and downlink bitrates perceived by the user at a particular time.

With respect to **MEC computation delay**, the task completion times depend on where the computation is performed. If it is done locally, that is, in the local user CPU, only the computation time is taken into account. However, if the task scheduler decides to offload the computation, both the computation time and transmission times are taken into account. The computation time depends on both the number of operations needed (${}^{\text{Size}}_{\text{MEC}}$) and the server computation capacity (${}^{\text{Server}}_{\text{MEC}}$) in cycles/second. Consequently, the final task completion time (or response time) is

$$ {}_{\text{MEC}} \qquad \begin{cases} \frac{{}^{\text{Size}}_{\text{MEC}}}{{}^{\text{Server}}_{\text{MEC}}} \quad \frac{{}^{\text{UL}}_{\text{MEC}}}{{}_{\text{UL}}} \quad \frac{{}^{\text{DL}}_{\text{MEC}}}{{}_{\text{DL}}} & \text{if offloading} \\ \frac{{}^{\text{Size}}_{\text{MEC}}}{{}^{\text{Server}}_{\text{MEC}}} & \text{otherwise} \end{cases} \tag{7}$$

Regarding the **C-RAN delay**, assuming that enough capacity is provided for the links connecting the RRUs and DUs, the queueing delay should be negligible. Then, the transmission delay of the fronthaul data between these two elements can be modeled as

$$ {}_{\text{RRU-DU}} \qquad \frac{{}^{\text{RU-DU}}_{\text{C-RAN}}}{{}_{\text{RU-DU}}} \tag{8}$$

where ${}^{\text{RU-DU}}_{\text{C-RAN}}$ and ${}_{\text{RU-DU}}$ are the size of the eCPRI burst and the channel capacity, respectively. Alternatively, if an aggregation network is used to merge the C-RAN traffic from several RRUs prior to sending the data to a DU, special attention must be put on the queueing delay. In that case, we refer the reader to [10] where we present an extensive study about the modeling of the queueing delay for eCPRI fronthaul flows merging in an Ethernet aggregator.

Once the C-RAN traffic enters the EPON, it is treated as any other type of traffic. Accordingly, the average packet delay in the upstream/donwstream directions of an EPON ($_{\text{PON}}$/ $_{\text{PON}}$) are taken from [24] as

$$^{\text{UL}}_{\text{PON}} \qquad _{\text{PON}} \frac{^{\text{UL}}}{^{\text{UL}}} \quad \frac{\overline{}}{_{\text{PON}}} \tag{9a}$$

$$^{\text{DL}}_{\text{PON}} \qquad _{\text{PON}} \quad \frac{\overline{}}{_{\text{PON}}} \tag{9b}$$

where $^{\text{UL}}$ is the occupancy of the EPON's uplink, $_{\text{PON}}$ is the propagation delay, $\overline{}$ is the average packet size, and $_{\text{PON}}$ is the uplink/downlink PON capacity.

Finally, the total response time for a user depends on the requested services as well as on the path of communication. Additionally, we assume for simplicity that the C-RAN delay budget (the worst-case budget for split processing) is subtracted from the service delay threshold. This means that, if a user accessing the network via C-RAN requests a service with a delay threshold of , the user will be treated by the system as a user requesting the same service but with a target delay threshold of $^{\text{Worst-case processing budget}}_{\text{C-RAN}}$. Also, note that, with this methodology, we make sure that the worst-case processing budget for C-RAN is met and taken into account by the optimization algorithm. Conversely, if we were only measuring the splits processing times, we would have no control over the delay requirement compliance.

F. Power Consumption Models

In order to compare all the operational approaches, we model the energy consumption to assess the optimality of each solution from an energy efficiency perspective. Mainly, two power consuming elements are taken into consideration: transmission and task execution. Regarding the communications, we model the consumption of an RF transmission in the uplink as

$$^{\text{UL}}_{\text{RF}} \qquad ^{\text{TX}} \qquad ^{\text{TX}} \quad _{\text{Device}} \quad ^{\text{UL}}_{\text{Comm}} \tag{10}$$

where $^{\text{TX}}$ is the static power consumption of the RF circuit, $^{\text{TX}}$ is the transmitter power consumption, which increases linearly with the emitted power ($_{\text{Device}}$), and $^{\text{UL}}_{\text{Comm}}$ is the transmission time of the data, computed as in Section III-E. Similarly, the downlink power consumption is modeled as

$$^{\text{DL}}_{\text{RF}} \qquad ^{\text{RX}} \qquad ^{\text{RX}} \quad _{\text{DL}} \quad ^{\text{DL}}_{\text{Comm}} \tag{11}$$

Note that here the transmitter power consumption depends on the downlink rate ($_{\text{DL}}$) rather than on the radiated power [25]. Adopting the work presented in [26], [27], we consider that the power consumption of an MU's computation task depends on either (a) the number of required CPU cycles ($^{\text{Size}}_{\text{MEC}}$) if the task is executed locally, or (b) the power needed to transmit the task to and receive the answer from a MEC server.

$$_{\text{MEC}} \begin{cases} ^{\text{UL}}_{\text{RF}} \quad ^{\text{DL}}_{\text{RF}} & \text{if offloading} \\ _{\text{local}} \quad ^{\text{Size}}_{\text{MEC}} & \text{otherwise} \end{cases} \tag{12}$$

where $_{\text{local}}$ represents the effective switched capacitance of the particular chip architecture and is the CPU's performance in number of operations per second.

IV. Network Utility Maximisation

In this Section, we model the DUDe scheme as a generalized *Network Utility Maximisation* (NUM) problem adapted to our architecture. Under a *Single Station Association* (SSA) policy, a mobile user is attached to a single base station in each link (downlink and uplink). Computing both an optimal *association* of MUs to base stations and optimal *allocation* of the BS resources to every MU is, under SSA policy, a well-known NP-hard combinatorial problem [28] whose complexity grows exponentially with the number of BS and MUs. We shall address this by adopting a *Multi-Station Association* (MSA) policy, allowing each MU to use multiple BS per link and simplify the solution later.

A. The Mixed NUM Problem

Let $_{\text{DL}}$ be a set of BSs capable of providing a downlink service, $_{\text{UL}}$ the set of BSs capable of providing an uplink service, and the finite set of users. The *maximum achievable rate* for user in the uplink (+) or downlink (-) direction with respect to the serving BS is the ergodic (Shannon) channel capacity given by

$$\text{BW}^{\text{DL}} \qquad \mathsf{SINR} \tag{13}$$

$$\text{BW}^{\text{UL}} \qquad \mathsf{SINR} \tag{14}$$

where the signal-to-interference-plus-noise ratio in the downlink for a user is given by Eq. (3), SINR is defined similarly for the uplink, and BW^{DL}/BW^{UL} are the uplink/downlink channel bandwidths [29], [30].

Let , a real non-negative variable, be the fraction of resources that BS grants to MU , representing a fraction of the total bandwidth. Let us normalize the maximum amount of resources that a BS can allocate to one, so that and . The feasible allocations for each link are the closed and convex sets

$$\mathbb{R} \quad ^{\text{DL}} \qquad _{\text{DL}}$$

$$\mathbb{R} \quad ^{\text{UL}} \qquad _{\text{UL}}$$

Note that the resource allocation variables can be used to indicate if a user is associated to a given base station, that is, a user is associated to BS in uplink if and it is not, otherwise. Then, the sum-rates for downlink and uplink for user are, respectively,

$$\tag{15}$$
$$_{\text{DL}}$$

$$\tag{16}$$
$$_{\text{UL}}$$

The objective is to maximize, for each user, a utility function that is continuously differentiable, monotonically increasing and strictly concave. These conditions hold for most of the utility functions considered in the literature and, in particular, for the class of α-proportional fair utility functions [31], [32], defined as follows

$$U_\alpha(x) = \begin{cases} \frac{x^{1\ \ \alpha}}{1-\alpha}, & \alpha \geq 0,\, \alpha \neq 1 \\ \log(x), & \alpha = 1. \end{cases} \tag{17}$$

The special case $\alpha = 0$ gives linear utility, $\alpha = 1$ reduces to logarithmic utility, and $\alpha \to \infty$ is equivalent to max-min fairness. Accordingly, the objective function becomes

$$f_\alpha \equiv \max_{\mathbf{y}\ \ \in \mathcal{Y}_{\text{DL}}, \mathbf{y}^+ \in \mathcal{Y}_{\text{UL}}} \sum_u [U_\alpha(\sum_{b_{\text{DL}}} r_{ub}^- y_{ub}^-) + U_\alpha(\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+)]. \tag{18}$$

Assuming that the uplink and downlink of the EPON are handled separately, Eq. 18 clearly decouples between uplink and downlink parts and can be solved separately. Knowing that the uplink poses more stringent delay requirements than the downlink (e.g., C-RAN uplink delay requirement, large MEC uplink computation packets and short downlink packets, ...), we shall provide the solution for the uplink and suggest the reader to apply the same methodology for the downlink.

The goal is then to achieve the optimal association of users to base stations and the optimal resource allocation to each user so that the utility is maximized for all users while the delay restrictions of the different services are met. Consequently, taking into account the constraints on $\mathbf{y}_u^+$ and the uplink end-to-end delay requirements of Section III-D, we can formulate the canonical optimization problem for the uplink as:

$$f_\alpha^{\text{Uplink}} \equiv \max_{\mathbf{y}^+ \in \mathcal{Y}_{\text{UL}}} \sum_u [U_\alpha(\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+)] \tag{19}$$

such that

$$\frac{S_{\text{H2H}}^{\text{UL}}}{\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+} + x_u^{\text{backhaul}} D_{\text{PON}}^{\text{UL}} \leq D_{\text{H2H}}^{\text{Threshold}}, \forall\, u \in \mathcal{U} \tag{20a}$$

$$\frac{S_{\text{H2M}}^{\text{UL}}}{\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+} + x_u^{\text{backhaul}} D_{\text{PON}}^{\text{UL}} \leq D_{\text{H2M}}^{\text{Threshold}}, \forall\, u \in \mathcal{U} \tag{20b}$$

$$\begin{aligned} \sum_{b_{\text{UL}}} \Bigg[& x_{ub}^{\text{MEC(Edge)}} \left(\frac{O_{\text{MEC}}^{\text{Size}}}{C_{\text{Server(b)}}^{\text{Edge}}} + \frac{S_{\text{MEC}}^{\text{UL}}}{r_{ub}^+ y_{ub}^+} \right) + \\ & + x_{ub}^{\text{MEC(CO)}} \left(\frac{O_{\text{MEC}}^{\text{Size}}}{C_{\text{Server}}^{\text{CO}}} + \frac{S_{\text{MEC}}^{\text{UL}}}{r_{ub}^+ y_{ub}^+} + \frac{D_{\text{PON}}^{\text{UL}}}{x_{ub}^{\text{MEC(CO)}}} \right) + \\ & + x_{ub}^{\text{MEC(Local)}} \left(\frac{O_{\text{MEC}}^{\text{Size}}}{C_{\text{Server}}^{\text{Local}}} \right) \Bigg] \leq D_{\text{MEC}}^{\text{Threshold}}, \forall\, u \in \mathcal{U} \end{aligned} \tag{20c}$$

$$\sum_{b_{\text{UL}}} (x_{ub}^{\text{MEC(Edge)}} + x_{ub}^{\text{MEC(CO)}} + x_{ub}^{\text{MEC(Local)}}) \leq 1, x_{ub} \in \mathbb{R}_+, \forall\, u \in \mathcal{U} \tag{20d}$$

$$\sum_{u \in \mathcal{U}} y_{ub}^+ \leq 1, \forall\, b \in \mathcal{B}_{\text{UL}}, \mathbf{y}^+ \in \mathcal{Y}_{\text{UL}}, \tag{20e}$$

where $y_{ub}^+ \geq 0$. Constraints (20a), (20b), and (20c) establish a maximum delay per user for each service. Also, constraints (20d) and (20e) are normalization constraints enforcing that we divide the tasks properly and that base stations do not grant more resources than those available, respectively. Note that x_u^{backhaul} is active if user u is using the backhaul for that particular communication process. Variables $x_{ub}^{\text{MEC(Edge)}}$, $x_{ub}^{\text{MEC(CO)}}$, and $x_{ub}^{\text{MEC(Local)}}$ account for the fraction of the task that is going to be executed at the edge, CO, and locally, respectively.

Lemma 1. *Choose feasible allocation schemes $\mathbf{y}^-$, $\mathbf{y}^+$ for downlink, uplink. Then, problems* (18) *and* (19) *are convex.*

Proof. If $\mathbf{y}^-$ and $\mathbf{y}^+$ are feasible allocation schemes, the objective function is a sum of a composition of a concave function with affine functions of $\mathbf{y}^-$, $\mathbf{y}^+$. Therefore, the objective function is concave itself. The constraints (20a), (20b) and (20c) are easily seen to be concave upward functions for $y_{ub}^+ \geq 0$. Also, constraints (20d) and (20e) are affine functions of $\mathbf{x}_{ub}$ and $\mathbf{y}_u^+$. In addition, the feasible allocation sets $\mathcal{Y}_{\text{DL}}$ and $\mathcal{Y}_{\text{UL}}$ are easily seen to be convex. Note, however, that the utility of the aggregated rate used by a given user in the downlink or in the uplink is not strictly (or strongly) concave, since an equation of the form $\sum_b r_{ub} y_{ub} = C$ may have multiple solutions on y_{ub}. This will be solved during the solution phase. □

B. Decentralized Solution

In this subsection, we perform a dual decomposition of the problem so as to identify decomposable structures. Using Lagrange duality theory, we connect the original maximization (19) with the dual problem (21). The generalized Lagrangian of the uplink problem is defined as

$$\begin{aligned} L(\mathbf{y}^+, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{x}) = & \sum_u [U_\alpha(\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+)] \\ & - \sum_u \lambda_u \left(\frac{S_{\text{H2H}}^{\text{UL}}}{\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+} + x_u^{\text{backhaul}} D_{\text{PON}}^{\text{UL}} - D_{\text{H2H}}^{\text{Threshold}} \right) \\ & - \sum_u \mu_u \left(\frac{S_{\text{H2M}}^{\text{UL}}}{\sum_{b_{\text{UL}}} r_{ub}^+ y_{ub}^+} + x_u^{\text{backhaul}} D_{\text{PON}}^{\text{UL}} - D_{\text{H2M}}^{\text{Threshold}} \right) \\ & - \sum_u \nu_u \Bigg[\sum_{b_{\text{UL}}} \Bigg[x_{ub}^{\text{MEC(Edge)}} \left(\frac{O_{\text{MEC}}^{\text{Size}}}{C_{\text{Server(b)}}^{\text{Edge}}} + \frac{S_{\text{MEC}}^{\text{UL}}}{r_{ub}^+ y_{ub}^+} \right) \\ & + x_{ub}^{\text{MEC(CO)}} \left(\frac{O_{\text{MEC}}^{\text{Size}}}{C_{\text{Server}}^{\text{CO}}} + \frac{S_{\text{MEC}}^{\text{UL}}}{r_{ub}^+ y_{ub}^+} + \frac{D_{\text{PON}}^{\text{UL}}}{x_{ub}^{\text{MEC(CO)}}} \right) \\ & + x_{ub}^{\text{MEC(Local)}} \left(\frac{O_{\text{MEC}}^{\text{Size}}}{C_{\text{Server}}^{\text{LOCAL}}} \right) \Bigg] - D_{\text{MEC}}^{\text{Threshold}} \Bigg] \\ & - \sum_u \gamma_u \left[\sum_{b_{\text{UL}}} (x_{ub}^{\text{MEC(Edge)}} + x_{ub}^{\text{MEC(CO)}} + x_{ub}^{\text{MEC(Local)}}) - 1 \right] \\ & - \sum_{b_{\text{UL}}} \alpha_b (\sum_u y_{ub}^+ - 1). \end{aligned} \tag{21}$$

By reworking and grouping the summations over users and over base stations, the Lagrangian becomes

UL
UL H2H backhaul UL PON Threshold H2H
UL
UL H2M backhaul UL PON Threshold H2M
UL
MEC Edge Size MEC Edge Server(b) UL MEC
UL
MEC(CO) Size MEC CO Server UL MEC UL PON MEC(CO) (22)
MEC Local Size MEC LOCAL Server Threshold MEC
MEC Edge MEC(CO) MEC Local
UL
UL UL

Now, the problem clearly separates into two different sides. On the one hand, the dual decomposition consists of each user solving the -th Lagrangian for the given vector of multipliers

MEC -
UL
UL H2H backhaul UL PON Threshold H2H
UL
UL H2M backhaul UL PON Threshold H2M
UL
MEC Edge Size MEC Edge Server(b) UL MEC
UL
MEC(CO) Size MEC CO Server UL MEC UL PON MEC(CO)
MEC Local Size MEC LOCAL Server Threshold MEC
MEC Edge MEC(CO) MEC Local
UL
UL
(23)

where each user knows its own multipliers , , , and . Multipliers can be understood as a price for each uplink allocation so that the users can choose the most favorable one. Recall that a similar Lagrangian can be posed for the downlink. On the other hand, the *master dual problem* can be written as

subject to
(24)

where , i.e., the Lagrangian for user evaluated at the optimal point. By Lemma 1 we know that the problem stated in (19) is differentiable in its domain. Accordingly, we may apply a *gradient projection method* to solve (24). Partial derivatives of the dual function give us the expressions to update the user's multipliers:

Threshold H2H UL H2H UL backhaul UL PON (25a)

Threshold H2M UL H2M UL backhaul UL PON (25b)

Threshold MEC
MEC Edge Size MEC Edge Server(b) UL MEC
UL
MEC(CO) Size MEC CO Server UL MEC UL PON MEC(CO) (25c)
MEC Local Size MEC LOCAL Server

MEC Edge MEC(CO) MEC Local
UL
(25d)

where is a sufficiently small positive step size and denotes the iteration index. Likewise, BS multipliers are updated as

UL (26)

Finally, the solution for the user's subproblems can be characterized by computing the Karush-Kuhn-Tucker (KKT) conditions for the Lagrangian. Assume we choose proportional *fairness* , i.e., . Then the partial derivative of the Lagrangian with respect to the decision variable is

UL H2H UL H2M
UL UL
UL MEC MEC Edge MEC(CO) UL
(27)

However, it is not clear to solve it as may have multiple solutions on . To overcome this, assume that only one base station is going to be used per link and choose the one that maximizes (23) according to the updated multiplier values,

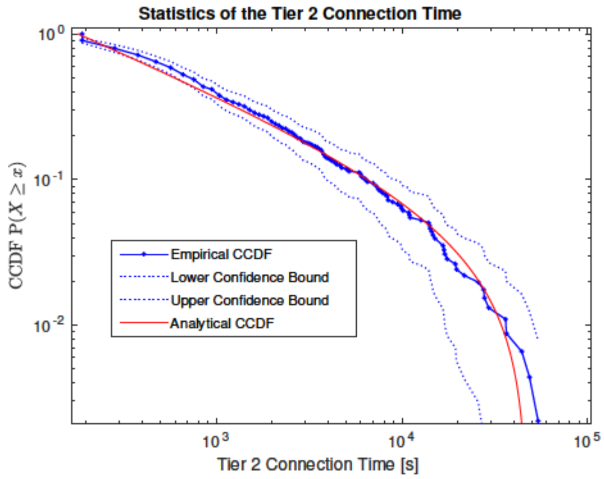


Fig. 2. CCDF of the tier 2 (WiFi or C-RAN) connection times.

removing the summations over b and assuming $y_{ub} = 1$. This requires, for each user, to check a number of base stations in range. As for decision of whether and where to offload a task, note that we can simplify the solution by choosing

$$[x_{ub}^{\text{MEC(Edge)}}, x_{ub}^{\text{MEC(CO)}}, x_{ub}^{\text{MEC(Local)}}] \in \{[1,0,0],[0,1,0],[0,0,1]\}, \tag{28}$$

adding 3 options per BS for MEC users. Confining the feasible solution set to (28) automatically satisfies constraint (20d) and removes the necessity of updating γ_u via (25d). Once that user u has chosen a base station b_i, the remaining task is to compute the optimal resource allocation y_{ub_i}. The partial derivative of the Lagrangian now simplifies to

$$\begin{aligned} \frac{\partial L_u}{\partial y^+_{ub_i}} &= \frac{r^+_{ub_i}}{r^+_{ub_i} y^+_{ub_i}} + \frac{\lambda_u S^{\text{UL}}_{\text{H2H}} + \mu_u S^{\text{UL}}_{\text{H2M}}}{r^+_{ub_i} {y^+_{ub_i}}^2} \\ &+ \frac{\nu_u S^{\text{UL}}_{\text{MEC}} (x^{\text{MEC(Edge)}}_{ub_i} + x^{\text{MEC(CO)}}_{ub_i})}{r^+_{ub_i} {y^+_{ub_i}}^2} - \alpha_{b_i} = 0. \end{aligned} \tag{29}$$

Solving for y_{ub_i}, we get the following polynomial:

$$-\alpha_{b_i} y^2_{ub_i} + y_{ub_i} + C = 0, \tag{30}$$

where the constant term is defined as

$$C = \frac{\lambda_u S^{\text{UL}}_{\text{H2H}} + \mu_u S^{\text{UL}}_{\text{H2M}} + \nu_u S^{\text{UL}}_{\text{MEC}} (x^{\text{MEC(Edge)}}_{ub_i} + x^{\text{MEC(CO)}}_{ub_i})}{r^+_{ub_i}}. \tag{31}$$

It is worth highlighting that the terms of (31) are only active when a user has any H2H, H2M or MEC request. Otherwise, note that $C = 0$ and the optimal resource allocation becomes pure logarithmic, as expected when choosing $\alpha-$*fairness* $= 1$.

V. Numerical Results

In this section, we validate our simulator and mobility model and compare the proposed solution with the baseline scenario.

TABLE I
Network deployment parameters.

Network Parameters	
Area of interest	1000m x 1000 m
Network deployment	LTE-A BS = 3 WiFi Spots = 9, RRUs = 6
Number of users	{25, 50, 75, 100, 125}
Transmit powers [16], [19]	LTE Base Stations = 46 dBm WiFi / C-RAN = 20 dBm
Path-loss exponent	4
Noise level	-106 dBm
Mobility model	Mobility model of [24]
C_{EPON}, $C_{\text{RU-DU}}$	25 Gb/s, 10 Gb/s
Background traffic ($R^{\text{UL}}_{\text{Wired}}$, $R^{\text{DL}}_{\text{Wired}}$)	5 Gbit/s, 7 Gbit/s [37].
Packet sizes	$S_{\text{H2H}} \sim 1,500$ B, $\sigma^2 = 0$ [24] $S_{\text{H2M}} \sim 88$ B, $\sigma^2 = 0$ (6-DoF [21]) $S^{\text{UL}}_{\text{MEC}} \sim \mathcal{N}(150\,\text{KB}, 100)$, $S^{\text{DL}}_{\text{MEC}} \sim \mathcal{N}(1,5\,\text{KB}, 10)$ [38]
MEC O_{size}	$\mathcal{U}(100, 200)$ MCycles [39]
MEC Servers Capacity (Consistent with [38])	CO = 3000 Mcycles/s Edge (ONU) = 2000 Mcycles/s Local (MU)= 1000 Mcycles/s
Energy Parameters [25], [27]	
k_1^{TX}, k_1^{RX}	0.4 W, 0.4 W
k_2^{TX}, k_2^{RX}	18, 2.86/10^6 W/bit
k_{local}, fi	10 ^{26}, 1000 cycles/s
Fixed Coverage Scenario	
P_{Offload}, (P_{Edge}, if offload)	0.5
Resource allocation	Pure Uniform: $y_{ub} = 1/\text{BS User}$
Optimization Parameters	
θ_{Users}	1.0
$\theta_{\text{BaseStation}}$	0.004

A. Validation of the Simulator and Mobility Model

The authors of [24] make use of the *PoneLab* [33] data to validate their analytical study that proposes an analytical expression for the CCDF of the WiFi connection times. Accordingly, we validate our simulator and mobility models by comparing this analytical model with our simulator. It consists of a truncated Pareto with the following parameters, taken from experimental measurements [34]: $\alpha_{\text{on}} = 0.54$, $\nu_{\text{on}} = 13.2$ hours, $\gamma_{\text{on}} = 3$ minutes, where ν_{on} and γ_{on} are the upper and lower bounds of the WiFi (tier 2) connection time, respectively.

Figure 2 plots the empirical CCDF of the WiFi/C-RAN connection times in our 2-tier HetNet and the truncated Pareto characterized by the previously mentioned parameters. Close inspection of the plots reveals that there is a good match between the experimental and the fitted distribution.

On the one hand, 99% confidence intervals are plotted for the empirical CCDF. Note how the the analytical CCDF of the truncated Pareto lies inside the confidence interval of the empirical CCDF. On the other hand, a Two-sample Kolmogorov-Smirnov test has been performed to asses if the two sample populations (the empirical and the analytical) might come from the same distribution. The results confirm that the null hypothesis H_0, that both samples come from populations with the same distribution, is not rejected for a significance level $\alpha_{\text{KS}} = 1\%$ with a *p-value* $= 0.0128 > \alpha_{\text{KS}}$ and a Kolmogorov-Smirnov statistic of $D = 0.1014$. Finally, the correlation between the empirical and analytical curves is 0.9009.

B. Use case: Tactile Internet

We now compare a Voronoi-based association and uniform resource allocation with the proposed optimization scheme. To that end, we set a common test scenario using the parameters settings in Table I. For this Tactile Internet use case, we

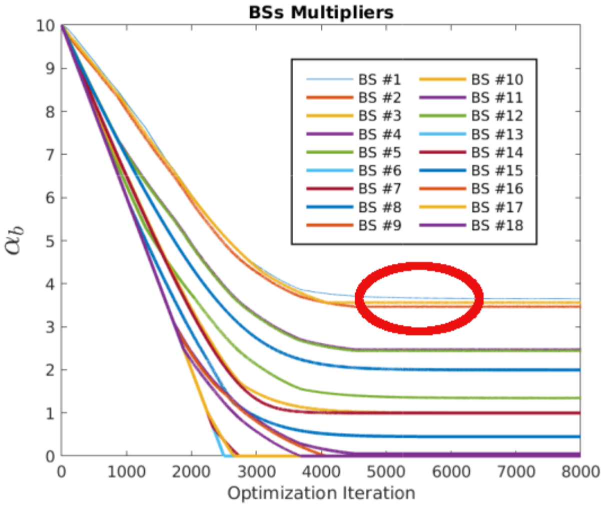


Fig. 3. Convergence of the multipliers announced by the base stations.

study (among other aspects), the successful integration of H2M remote control of haptic machines in a network characterized for a wide variety of technologies and services, as discussed in the introduction. The success of such services (and any other service) in these networks will be given by the ability to meet stringent latency requirements.

Regarding delay thresholds, $^{\text{Threshold}}_{\text{H2H}}$ ms (ITU-T Recommendation G.114: ms [35]) and $^{\text{Threshold}}_{\text{H2M}}$ ms are selected. MEC threshold is set to $^{\text{Threshold}}_{\text{MEC}}$ ms, consistent with the range in [36], so that we can assess the system's behavior when a solution is not always feasible. Regarding the offloading probabilities, $_{\text{offload}}$ $_{\text{edge}}$ are chosen for the baseline Voronoi-based scenario so that we have a representative sample of all the possible combinations (i.e., local computation, edge offloading, and CO offloading).

For the comparison of both approaches we assume that, at the beginning of every time-slot (s) of the simulation, all users have a request of a resource (i.e., transmission of H2H packet, H2M 6-DoF packet, MEC uplink and response packets). The one-way-delay of all requests are recorded. Also, the density and ratio ( ) between the number of small base stations (WiFi, RRU) and macro base stations (LTE-A) has been chosen to favor the uplink and downlink decoupling decisions in both scenarios, according to [20]. Channel bandwidth is MHz for C-RAN [10] and MHz for LTE-A and WiFi [16]. In addition, note that each one of the 6 C-RAN RRUs, under full utilization, is contributing with Mb/s to the EPON uplink [14]. Adding the fixed user traffic, this represents a total of Gb/s out of Gb/s utilization. This way, we test the network performance in an interesting high utilization regime where not all the decisions will comply with the delay requirements.

After enough number of iterations at each point in time, all multipliers converge to stable values, as shown in Figure 3 for . These multipliers reflect the state and needs of the network at every particular point in time. Then, these will auto adjust to changes, and cells will broadcast them so that they become part of users' computations. Once that users receive the updated values, they can choose the optimal base station by solving (23) and (28). Also, note that we take into account the network usage at each base station. Each BS multiplier is updated regarding the amount of resources left via (26). This causes that those base stations with a high occupancy or network usage start broadcasting a multiplier with a high value (see red circle in Figure 3), which suggests the users to choose another base station to access the network.

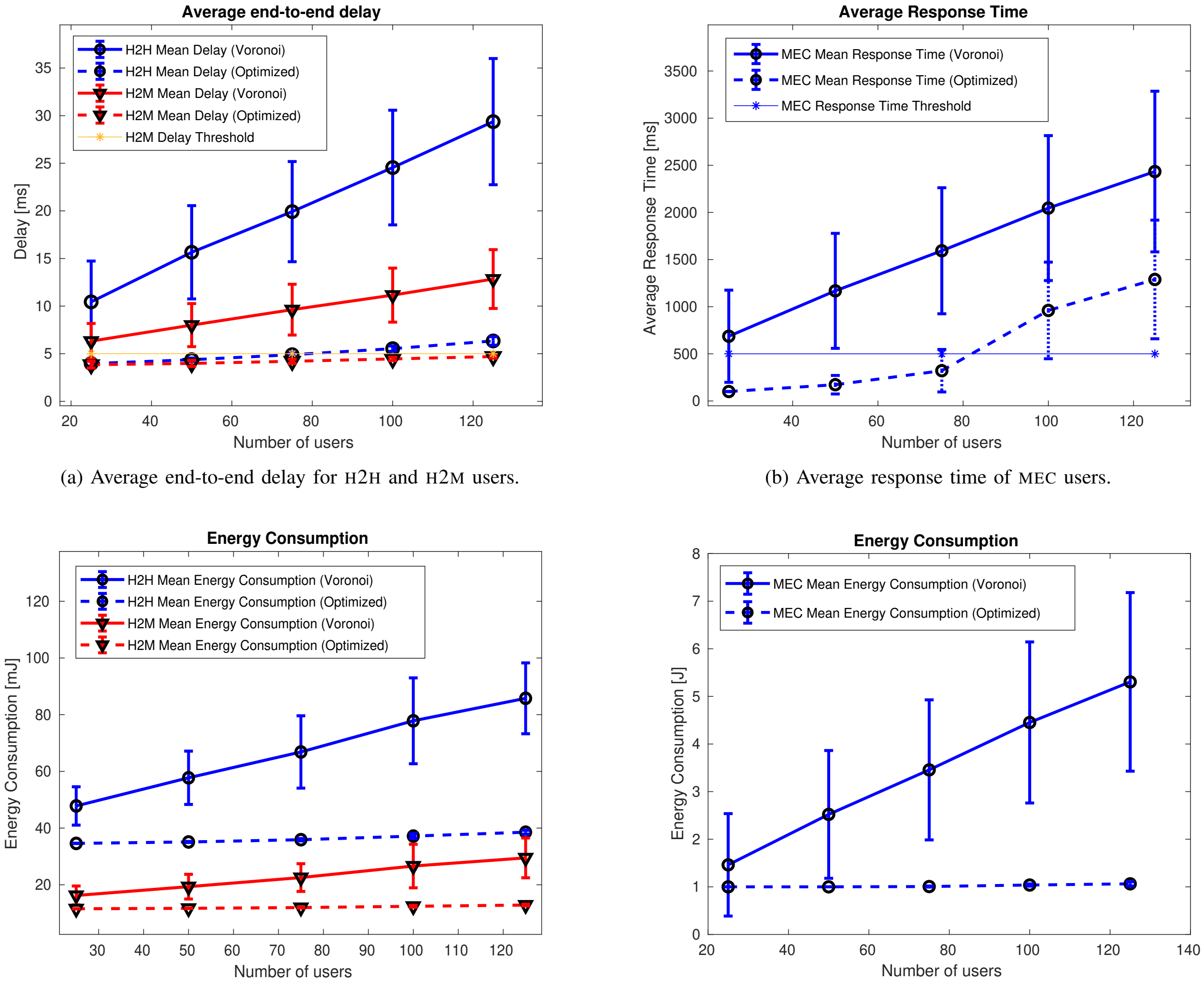
Figure 4a shows the evolution of the average end-to-end delay for H2H and H2M traffic classes, and Figure 4b illustrates the average response time for the MEC users. Recall that we use *response time* to denote the combination of both the task's communication end-to-end delay and computation time. Every point in the graph is averaged over 100 realizations, as we gradually increase the number of users. It is worth highlighting how the optimized solution outperforms the baseline solution in all cases.

Figure 4a (as well as the remaining ones), includes 95% confidence intervals for both the coverage/Voronoi-based and the optimization solutions. Note that the confidence intervals for the optimized solutions (see dashed lines) are so small and almost negligible. Observe that the delay threshold for the H2H users (ms) is met for both the Voronoi-based and the optimization solutions, achieving an average delay below the threshold, including the confidence interval. However, note that the optimization solution (see dashed blue line) achieves an average delay 5 times lower, in the worst-case utilization scenario (users). This is roughly an 80% improvement.

With regard to the H2M users, the traditional Voronoi approach is unable to meet their traffic delay threshold (set to ms). Conversely, the optimization approach is able to confine the average end-to-end delay always under ms. This holds true for all system loads. Taking a look at the confidence intervals, we conclude that the optimization solution is more stable, enabling us to ensure narrower and more precise delay intervals.

Close inspection of Figure 4b reveals that the optimization algorithm provides a better average response time for MEC users. For 75 users the decentralized algorithm achieves an average response time 5 times lower than that of the original approach. However, once that we exceed 80 users in the system, even the optimization is unable to find a suitable solution. Here, we find an interesting behavior of the proposed solution which is that, whenever a global solution is unfeasible, the service with the least stringent delay threshold is penalized.

Intuitively, we believe that it is a good choice to penalize those services with a larger delay threshold. The reason is that small deviations in the experienced service time will represent a smaller fraction of the expected delay threshold for these services, than it would for low delay threshold services. This behavior can be explained by inspection of the optimization's equations. In terms of the cost optimization of the objective function, it is more costly to leave unsatisfied a user with



(a) Average end-to-end delay for H2H and H2M users.

(b) Average response time of MEC users.

(c) Average H2H and H2M energy consumption.

(d) Average MEC energy consumption.

Fig. 4. Simulation results for the average one-way-delay and user energy consumption.

a low-latency requirement than to do it with a high latency threshold user. As we already mentioned, small deviations in the final service times for the low-latency threshold users, will rapidly increase the value of their multipliers (see (25a),(25b)). This gives them preference over those with a relaxed latency threshold in the resource allocation process.

Even though we think it is a desirable feature, we suggest several options to alter this behavior, if needed: (1) introduce additional priorities/fairness parameters to control preferences between small and big delay thresholds, (2) add supplementary restrictions to control the tightness of the thresholds, and (3) implement an adaptive convergence step for each user in the gradient descent algorithm, proportional to the size of its delay threshold.

Regarding possible scalability issues arising from this behavior: firstly, we think that it is a good feature that, even though not all of the users can be satisfied, the algorithm is still able to find the optimal solution with the available resources, so no user is left unattended. Secondly, this can be used as a warning flag to perform dynamic resource provisioning. If the number of processing units (MEC servers, C-RAN DUs, C-RAN CU pool, number of active femtocells, etc) is not enough, these can be dynamically instantiated or turned on. The same idea works the other way around, with the aim of saving resources. It is worth highlighting that there is no need for a protocol to include new users or resources into the optimization. Since the algorithm is decentralized, all parameters are available to the participants by just sensing the channel. Once that some new (virtual or physical) resources are instantiated or installed, the decentralized algorithm converges to the new global solution seamlessly. We believe that overall, these features are a good asset to achieve scalability.

Finally, Figures 4c and 4d show the H2H, H2M, and MEC user's energy consumption. Again, in Figure 4c the optimization-based solution is able to reduce the energy consumption for all kinds of services, spanning, for H2H, from

30% (with 25 users) to 50% (with 125 users). Regarding H2M services, we find power consumption improvements ranging from 28% to 56%, for 25 users and 125 users, respectively. Also, note that the H2M range is wider and shows a better improvement when the system is loaded, probably due to the fact that these users are favored during the optimization process. It is worth saying that the power consumption curves are also monotonically increasing in the optimized case. However, since the delay for these services is confined to lower values even with a high system occupancy (see Figure 4a), the increment rate is much lower than that of the Voronoi-based case. Likewise, the narrower confidence intervals for the optimized solution, suggest a higher stability and reliability of the system.

In Figure 4d, as the number of users increases, so does the network load and overall latency for MEC users. Due to the fact that in the Voronoi-based approach the MEC users are limited by fixed offloading decisions based on coverage maps, they keep offloading the tasks even though this is no longer the optimal decision. As the network delay increases, they must turn their radios on more time than before, causing the increase in the power consumption. Conversely, in the optimization-based solution, users change to local computation as soon as the alternative is no longer worthwhile and keep their power consumption stable.

After carefully reviewing the simulation results, it must be stressed that the optimization of the resource allocation and user association improves the overall throughput. Consequently, the delay for H2H, H2M and MEC services is reduced. Higher throughput also leads to shorter transmission times of a given data load, which can be translated into a reduced energy consumption.

VI. Conclusion

This paper considers the problem of decentralized resource allocation and user association in FiWi enhanced H-CRAN networks, for the converged technologies to support post-5G/6G services.

One of the concerns while merging all the envisioned technologies is to provide a converged network infrastructure to support them, particularly if these traditionally use different architectures. Since we are combining diverse wireless and fixed technologies that traditionally use their own dedicated equipment and orchestration (i.e., WiFi, LTE-A, C-RAN), especial attention must be put to how the integration is made and how the orchestration is performed so as to meet the requirements of all these technologies.

In this paper, a FiWi network topology based on a 25 Gbps EPON is proposed. We formulate an optimization problem to maximize the utility of all users while meeting the delay requirements of different services, that can be solved using a distributed iterative algorithm. The distributed algorithm is computationally simple as it only requires broadcasting the Lagrange multipliers. These work as a sort of price indicator, transmitting information about the state and needs of base stations and users at any particular time. Simulations show that this solution outperforms the classical received-power criteria in terms of average delay, power consumption and delay thresholds compliance. Namely, the results suggest that the distributed approach can achieve up to an 80% improvement in terms of average delay. More importantly, it is able to organize the user association and network resources in order to comply with the delay requirements.

Simulations show that this is not the case when relying on the traditional coverage maps. As a side effect, the energy consumption reduction obtained using the proposed approach is ranges from 28-56%, for H2M traffic, to 30-50% for H2H traffic.

Finally, we believe that a critical research point for further improvements of the optimization algorithm is the influence of the -fairness parameter in the power consumption and, in general, the effects of other utility functions that can be defined for the users. In this paper, a pure logarithmic resource allocation ( ) was chosen. Other allocation flavors might produce different results regarding the power consumption since it is not directly included in the optimization as a restriction. Consequently, we think this is a promising line for future work. Other future lines include adding network usage constraints, such as a maximum network usage of a certain technology.

Acknowledgments

Gabriel Otero Pérez would like to acknowledge the support of the Spanish Ministry of Education, Culture and Sport by means of the FPU grant he holds (Grant No. FPU16/01760). Universidad Carlos III de Madrid would like to acknowledge the support of EU projects PASSION (grant no. 780326) and BlueSpace (Grant No. 762055), and the Spanish project TEXEO (grant no. TEC2016-80339-R). Martin Maier acknowledges the support of the U. Carlos III de Madrid - Banco de Santander Excellence Chair programme.

References

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

[2] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 6, pp. 573-581, June 2018.

[3] ITU-T Technical Report GSTR-TN5G, "Transport network support of IMT-2020/5G," Feb. 2018.

[4] C. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5G perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66-73, Feb. 2014.

[5] M. Ali, Q. Rabbani, M. Naeem, S. Qaisar, and F. Qamar, "Joint user association, power allocation, and throughput maximization in 5G H-CRAN networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9254-9262, Oct. 2017.

[6] Saad, Walid, Mehdi Bennis, and Mingzhe Chen. "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems, " *arXiv preprint*. arXiv:1902.10265, 2019.

[7] 6Genesis Project, University of Oulu, Finland [Online]: https://www.oulu.fi/6gflagship.

[8] IEEE P802.3ca 50G-EPON Task Force, "Physical layer specifications and management parameters for 25Gb/s and 50Gb/s Passive Optical Networks".

[9] ITU-T Recommendation Supplement 66 to ITU-T G-series Recommendation, “5G wireless fronthaul requirements in a passive optical network context,” Series G: Transmission Systems and Media, Digital Systems and Networks, Oct. 2018.

[10] G. Otero Pérez, D. Larrabeiti López, and J. A. Hernández, “5G new radio fronthaul network design for eCPRI-IEEE 802.1CM and extreme latency percentiles,” *IEEE Access*, vol. 7, pp. 82218-82230, 2019.

[11] Common Public Radio Interface: Requirements for the eCPRI Transport Network. “eCPRI Transport Network V1.1”, 2017/10/24.

[12] IEEE Standard for Local and metropolitan area networks-IEEE Time-Sensitive Networking for Fronthaul IEEE Std 802.1CM.

[13] ETSI TS 138 470 V15.2.0 (2018), 5G; NG-RAN; “F1 general aspects and principles (3GPP, TS 38.470 version 15.2.0 Release 15)”.

[14] 3GPP TR 38.801, “Technical Specification Group Radio Access Network; Study on new radio access technology: Radio access architecture and interfaces, March 2017.

[15] A. Reznik, L. M. Contreras, and Y. Fang, “Cloud RAN and MEC: A Perfect pairing, ” ETSI White Paper No. 23, Feb. 2018.

[16] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, “Downlink and uplink decoupling: A disruptive architectural design for 5G networks,” in *Proc. IEEE GLOBECOM*, Austin, TX, 2014, pp. 1798-1803.

[17] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, “Optimal downlink and uplink user association in backhaul-limited HetNets,” in *Proc. IEEE INFOCOM 2016*, San Francisco, CA, 2016, pp. 1-9.

[18] F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski, and S. Singh, “Why to decouple the uplink and downlink in cellular networks and how to do it,” *IEEE Communications Magazine*, vol. 54, no. 3, pp. 110-117, March 2016.

[19] K. Smiljkovikj and H. Elshaer, “Capacity Analysis of Decoupled Downlink and Uplink Access in 5G Heterogeneous Systems,” *arXiv preprint* arXiv:1410.7270, 2014.

[20] K. Smiljkovikj, P. Popovski, and L. Gavrilovska, “Analysis of the Decoupled Access for Downlink and Uplink in Wireless Heterogeneous Networks,” *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 173-176, April 2015.

[21] M. Maier and A. Ebrahimzadeh, “Towards immersive Tactile Internet experiences: Low-latency FiWi enhanced mobile networks with edge intelligence [invited],” *IEEE/OSA Journal of Optical Communications and Networking, Special Issue on Latency in Edge Optical Networks*, vol. 11, no. 4, pp. B10-B25, April 2019.

[22] NTT DoCoMo Inc., “CU-DU split Refinement for Annex A (Transport network and RAN internal functional split),” TSG-RAN Working Group 3 meeting, R3-162102, October 2016.

[23] Small Cell Forum Document SCF159, “Small cell virtualization functional splits and use cases”, SCF159.05.1.01, June 2015.

[24] H. Beyranvand, M. Lévesque, M. Maier, J. A. Salehi, C. Verikoukis , and D. Tipper, “Toward 5G: FiWi enhanced LTE-A HetNets with reliable low-latency fiber backhaul sharing and WiFi offloading,” *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 690-707, Apr. 2017.

[25] O. Muñoz, A. Pascual-Iserte, and J. Vidal, “Optimization of Radio and Computational Resources for energy efficiency in latency-constrained application offloading,” *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4738-4755, Oct. 2015.

[26] Y. Xiao and M. Krunz, “QoE and power efficiency tradeoff for fog computing networks with fog node cooperation,” *IEEE INFOCOM*, Atlanta, GA, pp. 1-9, 2017

[27] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, “Mobile-edge Computing: partial computation offloading using dynamic voltage scaling,” *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268-4282, Oct. 2016.

[28] Z. Luo and S. Zhang, “Dynamic Spectrum Management: Complexity and Duality,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57-73, Feb. 2008.

[29] A. S. Cacciapuoti, “Mobility-Aware User Association for 5G mmWave Networks,” in *IEEE Access*, vol. 5, pp. 21497-21507, 2017.

[30] M. Elkourdi, A. Mazinm, and R. D. Gitlin, “Towards Low Latency in 5G HetNets: A Bayesian Cell Selection / User Association Approach,” 2018 *IEEE 5G World Forum (5GWF)*, Silicon Valley, CA, 2018, pp. 268-272.

[31] J. Mo and & J. Walrand, “Fair end-to-end window-based congestion control,” IEEE/ACM Transactions on networking, vol. 8, no. 5, pp. 556-567, Oct. 2000.

[32] M. Uchida and & J. Kurose, “An information-theoretic characterization of weighted alpha-proportional fairness,” in *Proc. IEEE INFOCOM*, 2009, pp. 1053-1061.

[33] A. Nandugudi et al., PhoneLab: A large programmable smartphone testbed, in Proc. *1st Int. Workshop Sens. Big Data Mining*, pp. 16, 2013.

[34] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, “Mobile data offloading: How much can WiFi deliver?” *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 536-550, Apr. 2013.

[35] Recommendation ITU-T G.114, “One-Way Transmission Time, Intl Telecommunication Union, ”Geneva (1996).

[36] H. Guo and J. Liu, “Collaborative computation offloading for multiaccess edge computing over fiber-wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4514-4526, May 2018.

[37] S. Mondal, G. Das, and E. Wong, “CCOMPASSION: A Hybrid Cloudlet Placement Framework Over Passive Optical Access Networks,” in *Proc. IEEE INFOCOM*, Honolulu, HI, 2018, pp. 216-224.

[38] S. Guo, B. Xiao, Y. Yang, and Y. Yang, “Energy-efficient dynamic offloading and resource scheduling in mobile cloud computing,” in *Proc. IEEE INFOCOM*, San Francisco, CA, 2016, pp. 1-9.

[39] S. Sundar and B. Liang, “Offloading dependent tasks with communication delay and deadline constraint,” in *Proc. IEEE INFOCOM*, Honolulu, HI, 2018, pp. 37-45.

[40] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and Andrews, J. G., “User association for load balancing in heterogeneous cellular networks, ” *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706-2716, June 2013.

[41] D. Pradas and M. A. Vazquez-Castro, “NUM-based fair rate-delay balancing for layered video multicasting over adaptive satellite networks,” *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 5, pp. 969-978, May 2011.

[42] K. Ronasi, A. Mohsenian-Rad, V. W. S. Wong, S. Gopalakrishnan, and R. Schober, “Delay-throughput enhancement in wireless networks with multipath routing and channel coding,” *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 1116-1123, March 2011.

[43] A. Sang, X. Wang, M. Madihian, and R.D. Gitlin, “A load-aware handoff and cell-site selection scheme in multi-cell packet data systems, ” *IEEE GLOBECOM*, volume 6, December 2004.

[44] R. Sivaraj, I. Broustis, N. K. Shankaranarayanan, V. Aggarwal, R. Jana, and P. Mohapatra, “A QoS-enabled holistic optimization framework for LTE-Advanced heterogeneous networks,” in *Proc. IEEE INFOCOM*, San Francisco, CA, 2016, pp. 1-9.

[45] F. Zheng, W. Li, P. Yu, and L. Meng, “User Association Based Cooperative Energy-Saving Mechanism in Heterogeneous 5G Access Networks,” in *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Chengdu, 2016, pp. 765-768.

[46] D. Liu, L. Wang, Y. Chue, M. Elkashlan, K. Wong, R. Schober, and L. Hanzo, “User Association in 5G Networks: A Survey and an Outlook,” in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1018-1044, Secondquarter 2016.

[47] S. Kassir, G. d. Veciana, N. Wang, X. Wang, and P. Palacharla, “Enhancing cellular performance via vehicular-based opportunistic relaying and load balancing,” in *Proc. IEEE INFOCOM*, Paris, France, 2019.

[48] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, “Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks,” *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275-5287, Nov. 2015.

[49] A. Alabbasi, M. Berg, and C. Cavdar, “Delay constrained hybrid CRAN: A functional split optimization framework,” in *Proc. IEEE GLOBECOM Workshops*, Abu Dhabi, United Arab Emirates, pp. 1-7, 2018.

[50] Y. Luo, F. Effenberger, and M. Sui, “Cloud computing provisioning over passive optical networks,” in *Proc. IEEE ICCC*, China, 2012, pp. 255-259.

[51] A. S. Reaz, V. Ramamurthi, M. Tornatore, and B. Mukherjee, “Cloud-integrated WOBAN: An offloading-enabled architecture for service-oriented access networks,” *Computer Netw.*, vol. 68, pp. 5-19, Aug. 2014.

[52] G. Castellano, F. Esposito, and F. Risso, “A distributed orchestration algorithm for edge computing resources with guarantees,” in *Proc. IEEE INFOCOM*, Paris, France, 2019.