Energy-Efficient Multi-Stream Carrier Aggregation for Heterogeneous Networks in 5G Cellular Systems

Elias Chavarria-Reyes, *Member, IEEE,* and Ian F. Akyildiz, *Fellow, IEEE,* and Etimad Fadel.

Abstract—Multi-stream carrier aggregation (MSCA) has been recently proposed as a mechanism to increase the amount of bandwidth available to users for heterogeneous networks (HetNets) in 5G cellular systems. Previous studies have focused only on maximizing the network capacity and fairness, without considering the energy efficiency of the MSCA. In this paper, the use of MSCA to minimize the energy consumption in a multi-layer HetNet is studied. The convexity of the energy minimization problem is examined, leading to the need of a quasiconvex relaxation. With this approximation, an algorithm (BIMEM) is designed to solve the energy minimization problem and obtain an optimum cell-association policy. Since the operators are generally interested in a balance between the energy minimization and capacity maximization, such multi-objective optimization is needed, and we studied it in this paper. The two aforementioned conflicting objectives can be jointly analyzed and solved through scalarization, even though the energy minimization has a quasiconvex objective function, and not a convex one. Performance evaluation is provided to identify the achievable energy savings of our proposed algorithm and to characterize the trade-offs between the energy minimization and capacity maximization in a multi-layer HetNet in 5G systems that support MSCA.

Index Terms—Heterogeneous networks, energy saving, cellular networks, cell-association, multi-stream, carrier aggregation.

I. INTRODUCTION

One of the most effective methods to improve the performance in cellular networks is to increase the amount of utilized bandwidth. Therefore, to meet the requirements of IMT-Advanced[1] [1], as well as those of the operators, the Long Term Evolution (LTE) Advanced (LTE-Advanced) considers the use of bandwidths of up to 100MHz in several frequency bands [2]. These bands are set by the International Telecommunication Union (ITU) and include the following [3]: 450-470MHz, 698-960MHz, 1710-2025MHz, 2110-2200MHz, 2300-2400MHz, 2500-2690MHz, 3400-3600MHz. For the fifth generation (5G) cellular systems, the use of even higher and wider frequency bands, such as millimeter wave ones, is a key enabling technology [4] [5]. LTE-Advanced tries to exploit as much as possible the flexibility of supporting multiple frequency bands through the use of carrier aggregation [6] [7].

Elias Chavarria-Reyes and Ian F. Akyildiz are with the Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA (e-mail: elias.chavarria@gatech.edu and ian@ece.gatech.edu).
Etimad Fadel and Ian F. Akyildiz are with the Computer Science Department at the King Abdulaziz University, Jeddah, Saudi Arabia (email: eafadel@kau.edu.sa).

[1]IMT-Advanced stands for international mobile telecommunications-advanced.

Carrier aggregation (CA) consists of grouping several component carriers (CC) to achieve wider transmission bandwidths. An LTE-Advanced device can aggregate up to five CCs, each of up to 20 MHz. With the largest configuration, this implies a total bandwidth of 100MHz. To support backward compatibility with LTE user equipment (UE), each of the CCs shall be configured as a typical LTE carrier. Therefore, any of the CCs used for CA should also be accessible to LTE UEs. Nevertheless, mechanisms, such as barring [8], already exist to prevent LTE UEs from camping on specific CCs. This way, operators have the flexibility of adjusting the characteristics of the CCs to support a mixture of LTE and LTE-Advanced devices.

To obtain the most benefit from CA, each base station (BS) should support the maximum number of CCs[2]. Nevertheless, in current 4G and even more so in future 5G systems, cellular networks are composed of a combination of small cells that provide enhanced coverage in targeted areas and macrocells that provide basic coverage, creating a HetNet [9] [10]. In most HetNet scenarios, not all BSs support the maximum CA configuration. This is mainly due to the hardware limitations, which require costly upgrades. Multi-stream CA (MSCA), also known as multi-flow CA, has been recently proposed as an alternative method to address this problem [11] [12] [13]. In MSCA, a UE is able to aggregate CCs belonging to multiple BSs, allowing it to achieve the maximum CA configuration, even if no BS can provide such configuration by itself.

In non-MSCA networks, the existing literature has looked at multiple ways of reducing the network energy consumption. In [14], the use of lean carriers with reduced signaling overhead is proposed. By reducing the signaling overhead, the BS can go into micro-sleep more frequently. The concept of adjusting the cell-association policies and, therefore, the load across BSs has also been proposed separately for energy minimization [15] and user fairness [16] [17] [18]. Furthermore, cooperation among BSs has been utilized to minimize the energy consumption by coordinating the scheduling and power control mechanisms [19] [20] and the on-off policies [21] [22]. Compared to the literature on non-MSCA, existing work on MSCA-enabled networks has focused on maximizing the network capacity [23] [24], achieving a target SINR [25] [26], and the analysis of the impact of biasing and the selection of frequency bands [27] [28]. However, even though reducing the energy consumption is a key design factor for 5G cellular systems [10] [29], almost no work has been done on analyzing

[2]The current version of the standard specifies the maximum configuration to be five CCs.
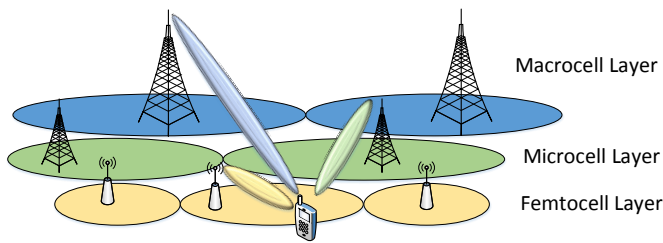
Fig. 1. MSCA in a HetNet.

energy-efficient methods of exploiting MSCA [30].

The focus of our work is designing new methods of exploiting MSCA to improve the energy efficiency in multi-layer HetNets for current 4G and future 5G cellular systems. In particular, we show that the energy minimization problem in MSCA-enabled networks is a non-convex optimization. Nevertheless, such problem can be approximated through a generalized linear-fractional program. Using this approximation, we develop a simple algorithm to solve such problem by applying a bisection method that solves a convex feasibility problem at each step, until a precision tolerance is met. Since the operators are typically interested not only in minimizing the energy consumption, but also in maximizing the network capacity, we analyze these problems jointly as a multi-objective optimization. Based on the analysis done for the energy minimization problem, we provide a solution for the multi-objective one, according to the priority assigned by the operators to each objective. Moreover, we show that an explicit analytical expression for the UE-to-CC association policy can be obtained without the need of solving the multi-objective optimization problem.

The rest of this paper is organized as follows. We present the network architecture and BS energy model in Sections II-A and II-B, respectively. In Section III, we develop the energy- and capacity-aware mechanisms of load balancing that exploit the use of MSCA. In particular, in Section III-A, we focus on the single objective of minimizing the network energy consumption. Then, the energy minimization and capacity maximization are jointly analyzed in Section III-B. Simulation results showing the performance of our load-balancing mechanisms are presented in Section IV. Finally, the conclusions are presented in Section V.

II. SYSTEM MODEL

A. Network Architecture

An example of a network where a UE and the BSs support MSCA is depicted in Figure 1. From Figure 1, we observe that the UE applies MSCA by connecting to three CCs that belong to BSs of different layers. This behavior follows from the fact that the BSs in the same layer are typically assigned the same frequency bands. Therefore, if a UE were to connect to multiple BSs of the same layer, it would have to utilize advanced intercell interference cancellation (ICIC) techniques to recover the signal from each BS [31] [32] [33]. While not impossible, such functionality is not within the scope of MSCA; rather, it is included in cooperative multipoint transmission and reception

(CoMP) [34] [35]. CoMP requires not only advanced ICIC at the UE, but also a significant amount of coordination between the BSs, so that their transmission scheme is suitable for the application of advanced ICIC at the UE [36] [37] [38]. Compared to CoMP, the overhead associated with MSCA is significantly smaller since there is no need to coordinate or synchronize among the CCs used for MSCA how the physical resources are utilized.

For the purpose of MSCA, we stipulate that any pair of BSs that belong to different layers will utilize CCs that operate on different frequency bands. On the other hand, BSs that belong to the same layer may utilize CCs that operate on the same frequency bands. We consider that the network serves a set of users $\mathcal{U}$ and that each user can generate sessions that have different quality of service (QoS) requirements $q \in \mathcal{Q}$. These are defined in terms of bit rates, to capture the non-linear relationship between bit rate and power. We also consider that each UE u_i has a set Γ_i of active CCs. A given Γ_i contains at most one CC for every frequency band. Among all the CCs that belong to the same frequency band, a given UE u_i will add to its Γ_i the CC towards which the value of pathloss plus shadowing is the smallest. While other methods of selecting the set Γ_i could be explored, they are out of the scope of this article and will be explored in future work. In addition, we consider that each UE is capable of partitioning every session across all of its active CCs. Such assumption is justified by the fact that, if we were to consider that a session can only be served through a single CC, the overall problem would be reduced to a flow-to-CC association problem, which is equivalent to the traditional UE-to-CC cell-association problem when no MSCA is considered. Moreover, we consider that the processes of partitioning the traffic at the UE and recombining such traffic at the network introduce a negligible bit rate overhead to the total bit rate required to satisfy a session. Furthermore, the analysis is done considering single-antenna systems; the extension to multi-antenna systems will be explored in future work. In terms of the channel, we consider path loss, shadowing and its auto-correlation effects, and non-selective block fading. From the network perspective, we consider the load of a BS to be represented by the total bit rate it provides to the UEs that it serves. Thus, from now on we will use the terms bit rate and load interchangeably. Our definition of network load represents the input that the network needs to support, regardless of how the network utilizes its own resources to handle the input - which is the definition of load used in some existing work [39]. As a result, with our definition, the load metric becomes a function of the traffic, i.e., the input, and not of what our algorithm does with the network resources.

B. Base Station Energy Model

In terms of the energy consumption of the BS, we consider that each BS b has a set $\mathcal{S}_b$ of CCs. From now on, we will utilize the notation $CC_{j,k}$ to denote CC k of BS j. We consider the energy consumption model of an active $CC_{j,k}$ during a time interval Δt to be

$$\hat{E}_{\text{total}}(CC_{j,k}) = \Delta t \left[\hat{P}_{\text{on,min}}(CC_{j,k}) + \hat{P}_{\text{on,dyn}}(CC_{j,k}) \right], \quad (1)$$

where $\hat{E}_{\text{total}}(\text{CC}_{j,k})$ represents the total energy consumption of $\text{CC}_{j,k}$, $\hat{P}_{\text{on,min}}(\text{CC}_{j,k})$ represents the amount of power consumed regardless of the traffic handled by $\text{CC}_{j,k}$, and $\hat{P}_{\text{on,dyn}}(\text{CC}_{j,k})$ represents the dynamic power consumption of $\text{CC}_{j,k}$, which varies with the traffic dynamics. Since $\hat{P}_{\text{on,min}}(\text{CC}_{j,k})$ is constant with respect to the traffic handled by a CC, it can be excluded from the rest of the analysis. We consider that $\hat{P}_{\text{on,dyn}}(\text{CC}_{j,k})$ is a linear function of the RF output power required to satisfy the QoS requirements requested by the UEs connected to that CC, as generally done in the literature [40] [41] [42]. Thus, we have that

$$\hat{P}_{\text{on,dyn}}(\text{CC}_{j,k}) = w_{j,k} \sum_i P_{i,j,k}, \qquad (2)$$

where $P_{i,j,k}$ represents the amount of RF output power required by $\text{CC}_{j,k}$ to satisfy the QoS requested by UE i and $w_{j,k}$ is a constant unique for every $\text{CC}_{j,k}$. The value of not only $w_{j,k}$, but also $\hat{P}_{\text{on,min}}(\text{CC}_{j,k})$, depends on the internal components associated with the operation of the CC and their interconnection [43] [44].

III. Energy- and Capacity-Aware Load Balancing

In a HetNet, the macrocells transmit at a much higher power than the small cells. Thus, regions will exist where a UE will be relatively close to a small cell, but the SINR of a macrocell will still be higher. Therefore, an association policy based on the CC that provides the maximum SINR, i.e., a max-SINR policy, tends to favor the association towards the macrocells over small cells in the aforementioned regions. As a result, a UE close to a small cell may still connect to a macrocell even if

- communicating with the macrocell requires more energy, in uplink or downlink, than communicating with the small cell or
- the macrocell is overloaded, and the small cell resources are being underutilized. Such situation may cause the user QoS requirements to not be satisfied by the macrocell, even though the small cell could have done so.

Moreover, even if a user is capable of following a cell-association policy different from the max-SINR one, the fact that it can only connect to a single BS means that

- no single cell may have enough capacity to satisfy the downlink and uplink QoS requirements or
- a user may connect to a cell that can satisfy the downlink or uplink QoS requirements, but not both or
- the non-linear relationship between the minimum power received and bit rate would require a disproportional amount of power to satisfy the QoS requirements.

Thus, MSCA can address the aforementioned issues, given that we design mechanisms to balance the load across small cells and macrocells while accounting for the energy consumption and the capacity of each one. The design of such mechanisms for the downlink is the focus of this work. The uplink could be analyzed by following a similar approach, assuming that an accurate model for the UE total energy consumption is available.

For UE i and $\text{CC}_{j,k}$, the maximum spectral efficiency $\theta_{i,j,k}$ of the communication link is a logarithmic function of the SINR:

$$\theta_{i,j,k} = \beta \log_2 \left(1 + \underbrace{h_{i,j,k} \frac{P_{i,j,k}}{\eta_{i,j,k}}}_{\text{SINR}} \right), \qquad (3)$$

where $P_{i,j,k}$ is the RF output power used by $\text{CC}_{j,k}$ on the resources assigned to UE i, $h_{i,j,k}$ is the channel gain between UE i and $\text{CC}_{j,k}$, $\eta_{i,j,k}$ represents the noise plus interference experienced by UE i when connected to $\text{CC}_{j,k}$, and $0 < \beta < 1$ is an attenuation factor that accounts for implementation losses and can be chosen to represent different modem implementations and link conditions [45]. The factor $h_{i,j,k}$ includes the path loss, fading, and shadowing effects. By considering a large time scale for the association between a user and a CC, the short-term channel dynamics, such as fast fading, can be averaged out, allowing us to consider the SINR and the spectral efficiency as constants during the association duration. For the factor $\eta_{i,j,k}$, we consider that there is no intra-cell interference and, to make the formulation tractable, that the expected inter-cell interference within a given layer is known or managed through existing techniques, such as inter-cell interference coordination.

Even though the maximum spectral efficiency $\theta_{i,j,k}$ is a good metric for the quality of the channel between user i and $\text{CC}_{j,k}$, the overall bit rate is the metric of interest when determining if the QoS is satisfied. The bit rate achieved over the aforementioned channel depends not only on $\theta_{i,j,k}$, but also on the amount of resources assigned to such channel by the BS. Particularly, for user i, $\text{CC}_{j,k}$ with bandwidth $\rho_{j,k}$, the maximum bit rate $\tilde{\theta}_{i,j,k}$ over a channel that has a maximum spectral efficiency $\theta_{i,j,k}$ is

$$\begin{aligned} \tilde{\theta}_{i,j,k} &= \rho_{j,k} y_{i,j,k} \theta_{i,j,k} \\ &= \rho_{j,k} y_{i,j,k} \beta \log_2 \left(1 + h_{i,j,k} \frac{P_{i,j,k}}{\eta_{i,j,k}} \right), \qquad (4) \end{aligned}$$

where the factor $0 \leq y_{i,j,k} \leq 1$ represents the fraction of resources reserved for user i by $\text{CC}_{j,k}$, and $\theta_{i,j,k}$ is obtained from Eq. (3). By considering that the resource allocation is performed within the coherence time of the channel, the latter can be considered static during every allocation period. Such assumption is valid for low-mobility scenarios.

Typically, $y_{i,j,k} < 1$ since the BS serves more than one user; therefore, it must allocate the limited resources among those users. As a result, the achievable rate of a user depends not only on the channel quality towards a particular CC, but also on the number of other users associated with such CC and the resource allocation policy followed. The latter depends directly on how much bit rate, i.e., load, each UE requests from each CC.

In Section III-A, we focus on finding an optimal load-balancing and cell-association policy that minimizes the energy consumption of the network. Then, in Section III-B, we utilize the results from Section III-A to develop an optimal load-balancing and cell-association policy capable

of addressing the conflicting objectives of minimizing the network energy consumption and maximizing its capacity.

A. Load Balancing for Energy Minimization

Based on the QoS requirements of all the sessions that a user i needs to support, a total bit rate r_i can be computed for such user. By using MSCA, r_i can be split across all the CCs to which the UE is capable of connecting. If UE i requests a fraction $0 \leq \xi_{i,j,k} \leq 1$ of r_i to $CC_{j,k}$, then the following relationship must hold:

$$\xi_{i,j,k} r_i \leq \tilde{\theta}_{i,j,k}$$
$$= \rho_{j,k} y_{i,j,k} \beta \log_2 \left(1 + h_{i,j,k} \frac{P_{i,j,k}}{\eta_{i,j,k}} \right), \quad (5)$$

where $\tilde{\theta}_{i,j,k}$ is obtained from Eq. (4). The above expression conveys that the amount of bit rate requested by any UE to any CC should not exceed the capacity of the channel between them for a given amount of bandwidth and power allocated to the user. Typically, the above inequality can be treated as an equality, since there is no benefit to the user or BS to have an underutilized channel. Assuming that $y_{i,j,k} > 0$, i.e., that $CC_{j,k}$ is assigning a non-zero amount of bandwidth to user i, the amount of output RF power at the antenna of $CC_{j,k}$ for user i can be obtained as follows,

$$P_{i,j,k} = \left[\exp \left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2) \right) - 1 \right] \frac{\eta_{i,j,k}}{h_{i,j,k}}$$
$$= \frac{\eta_{i,j,k}}{h_{i,j,k}} \exp \left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2) \right) - \frac{\eta_{i,j,k}}{h_{i,j,k}}. \quad (6)$$

As discussed in Section II-B, the total dynamic power $\hat{P}_{\text{on,dyn}}$ consumed by $CC_{j,k}$ to output the RF power required to satisfy the QoS requirements of the UEs is a linear function of $P_{i,j,k}, \forall i, j, k$:

$$\hat{P}_{\text{on,dyn}}(CC_{j,k}) = w_{j,k} \sum_i P_{i,j,k}$$
$$= w_{j,k} \sum_i \frac{\eta_{i,j,k}}{h_{i,j,k}} \exp \left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2) \right)$$
$$- w_{j,k} \sum_i \frac{\eta_{i,j,k}}{h_{i,j,k}}. \quad (7)$$

Based on the expression above, the network energy minimization problem can be described as

$$\text{minimize} \quad \sum_j \sum_k \hat{P}_{\text{on,dyn}}(CC_{j,k}), \quad (8a)$$

$$\text{subject to} \quad \sum_j \sum_k \xi_{i,j,k} = 1, \quad \forall i, \quad (8b)$$

$$\sum_i y_{i,j,k} \leq 1, \quad \forall j, k, \quad (8c)$$

$$\xi_{i,j,k} \geq 0, \quad \forall i, j, k, \quad (8d)$$

$$y_{i,j,k} \geq 0, \quad \forall i, j, k, \quad (8e)$$

$$r_i \xi_{i,j,k} - \rho_{j,k} \theta_{\max} y_{i,j,k} \leq 0, \quad \forall i, j, k, \quad (8f)$$

$$\xi_{i,j,k} = 0; \quad \forall i, j, k, CC_{j,k} \notin \Gamma_i \quad (8g)$$

$$y_{i,j,k} = 0; \quad \forall i, j, k, CC_{j,k} \notin \Gamma_i \quad (8h)$$

where $\theta_{\max}$ denotes the maximum spectral efficiency supported by the network[3]. Constraint (8b) indicates that the UE total QoS requirement r_i must be satisfied with equality. Constraint (8c) indicates that $CC_{j,k}$ cannot allocate more bandwidth than it has available. Constraints (8d) and (8e) indicate that the allocation variables $\xi_{i,j,k}$ and $y_{i,j,k}$ are non-negative. Constraint (8f) indicates that every channel should operate within the maximum spectral efficiency supported by the network. Constraints (8g) and (8h) indicate that a UE i can associate only with the CCs that belong to its active set Γ_i of CCs. All the constraints in the optimization problem (8) are linear expressions. Also, note that in addition to the intrinsic parameters associated with the CCs, the total QoS requirement of each UE and the SINR between the UEs and each BS CC would be required as inputs to solve the optimization problem (8).

The optimization problem (8) is equivalent to

$$\text{minimize} \quad \sum_{i,j,k} w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \exp \left(\frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2) \right), \quad (9a)$$

$$\text{subject to} \quad \text{Constraints (8b)-(8h)}, \quad (9b)$$

where the optimization variables are all the $\xi_{i,j,k}$ and $y_{i,j,k}$, and we drop the last term of Eq. (7) because it is a constant that does not affect the solution of the problem. It is important to highlight that objective function (9a) does not allow for $y_{i,j,k} = 0$, even though constraint (8e) does. We will later see that the objective function can be further approximated to allow $y_{i,j,k} = 0$.

Since the factors $w_{j,k}$, $\eta_{i,j,k}$, and $h_{i,j,k}$ are positive constants, the optimization problem (9) can be further rewritten as

$$\text{minimize} \quad \sum_{i,j,k} \exp \left(\alpha_{i,j,k} + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln(2) \right), \quad (10a)$$

$$\text{subject to} \quad \text{Constraints (8b)-(8h)}, \quad (10b)$$

where $\alpha_{i,j,k}$ is a constant defined as

$$\alpha_{i,j,k} \triangleq \ln \left(w_{j,k} \frac{\eta_{i,j,k}}{h_{i,j,k}} \right). \quad (11)$$

We now analyze the convexity of the optimization problem (10). Consider a single term of the summation in the objective function. Any such term is an exponential function whose argument is a linear-fractional function. Since a linear-fractional function is quasiconvex, and the exponential function is a non-decreasing function, it follows that each term in the above summation is also quasiconvex. Nevertheless, while convexity is preserved by a non-negative weighted sum operation, quasiconvexity may not be [46]. Therefore, the optimization problem (10) needs to be reformulated before any convex or quasiconvex optimization technique can be applied.

---

[3]In LTE, $\theta_{\max}$ is approximately 4.8 bits/sec/Hz, based on 64 QAM and 4/5 code rate [45].

First, since the logarithmic function is a monotonically increasing function, we apply it to the objective function (10a) to create an equivalent optimization problem:

$$minimize \quad \ln \left(\sum_{i,j,k} \exp \left(\alpha_{i,j,k} + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln (2) \right) \right),$$
(12a)

$$subject \ to \quad \text{Constraints (8b)-(8h)}. \tag{12b}$$

Second, we exploit the log-sum-exp approximation

$$\max_{i=1...n} x_i \leq \ln \sum_{x=1}^{n} e^{x_i} \leq \max_{i=1...n} x_i + \ln n. \tag{13}$$

The lower bound is met when there is only one non-zero x_i; the upper bound is met when all the x_i are equal. For a given n, minimizing $\max_{i=1...n} x_i$ reduces the values of both the upper and lower bounds of $\ln \sum_{x=1}^{n} e^{x_i}$. Applying the above approximation to the optimization problem (12), it then becomes

$$minimize \quad \max_{i,j,k} \left[\alpha_{i,j,k} + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln (2) \right], \tag{14a}$$

$$subject \ to \quad \text{Constraints (8b)-(8h)}. \tag{14b}$$

The objective function (14a) is a maximization of linear-fractional functions and, therefore, of quasiconvex functions. Such formulation is also known in the literature as a generalized linear-fractional program. Since a nonnegative weighted maximum of quasiconvex functions is also quasiconvex, so is the above objective function. We can now apply any general approach for quasiconvex programming. One such approach consists in representing the sublevel sets of the quasiconvex function via a family of convex inequalities [46], as we will now describe.

First, we define $g_0 (\xi, y)$ as our current objective function:

$$g_0 (\xi, y) = \max_{i,j,k} \left[\alpha_{i,j,k} + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln (2) \right]. \tag{15}$$

Second, for a given parameter μ, $g_0 (\xi, y) \leq \mu$ if and only if

$$\max_{i,j,k} [\xi_{i,j,k} r_i \ln (2) - \beta \rho_{j,k} y_{i,j,k} [\mu - \alpha_{i,j,k}]] \leq 0. \tag{16}$$

If we define the convex function $\vartheta_\mu (\xi, y)$ as

$$\vartheta_\mu (\xi, y) = \max_{i,j,k} [\xi_{i,j,k} r_i \ln (2) - \beta \rho_{j,k} y_{i,j,k} [\mu - \alpha_{i,j,k}]], \tag{17}$$

then

$$g_0 (\xi, y) \leq \mu \Leftrightarrow \vartheta_\mu (\xi, y) \leq 0. \tag{18}$$

Therefore, the μ-sublevel set of the quasiconvex function g_0 is the 0-sublevel set of the convex function ϑ_μ.

Let us denote the optimal value of the quasiconvex optimization problem (14) as χ^*. If the feasibility problem

$$find \qquad \xi, y, \tag{19a}$$

$$subject \ to \qquad \vartheta_\mu (\xi, y) \leq 0, \tag{19b}$$

$$\text{Constraints (8b)-(8h)}, \tag{19c}$$

is feasible, then $\chi^* \leq \mu$ and any feasible point (ξ, y) is also a feasible point for the quasiconvex problem (14). If the

problem (19) is not feasible, then $\chi^* > \mu$. Problem (19) is a convex feasibility problem. Therefore, we can verify whether χ^* is greater or less than a particular value μ by solving problem (19). Based on this last observation, a simple procedure to find χ^*, BIMEM, is designed through a bisection method that solves a convex feasibility problem at every step, as described in Algorithm 1.

Algorithm 1 BIMEM: Bisection method for energy minimization.

1: **given** $l_1 \leq \chi^*$, $l_2 \geq \chi^*$, $\epsilon > 0$
2: **repeat**
3:     $\mu = (l_2 + l_1) /2$
4:     Solve the convex feasibility problem (19)
5:     **if** (19) is feasible **then**
6:         $l_1 = \mu$
7:     **else**
8:         $l_2 = \mu$
9:     **end if**
10: **until** $l_2 - l_1 \leq \epsilon$

In Algorithm 1, assuming that the quasiconvex problem (14) is feasible and that we know an interval $[l_1, l_2]$ that contains the optimal value χ^*, we solve the feasibility problem at the midpoint $\mu = (l_1 + l_2)/2$ of such interval by applying any convex optimization technique, e.g., interior-point method. The result of the feasibility problem indicates whether χ^* is in the lower or upper half of the interval, which we then use to update the interval accordingly. The new interval is half the size of the initial one, i.e., it is bisected. This procedure is repeated until the size of the interval satisfies some lower bound ϵ. After m iterations, the size of the interval is $2^{-m}(l_2 - l_1)$. Therefore, the number of iterations required before the algorithm terminates is $\left\lceil \log_2 \left((l_2 - l_1)/\epsilon \right) \right\rceil$.

To apply Algorithm 1, we need the initial interval $[l_1, l_2]$ that is guaranteed to contain the optimal value χ^*. Such interval can be obtained from constraint (19b), as shown below. For such constraint to be satisfied, the following expression must be true:

$$\max_{i,j,k} [\xi_{i,j,k} r_i \ln (2) - \beta \rho_{j,k} y_{i,j,k} [\mu - \alpha_{i,j,k}]] \leq 0, \tag{20}$$

which is equivalent to

$$\xi_{i,j,k} r_i \ln (2) - \beta \rho_{j,k} y_{i,j,k} [\mu - \alpha_{i,j,k}] \leq 0, \qquad \forall i, j, k. \tag{21}$$

Since r_i is positive and $\xi_{i,j,k}$ is non-negative, then the first term is also non-negative. Therefore, we need the second term to be non-positive. As a result, there are two possible necessary conditions for the above expression to be satisfied for any given i,j,k:

$$\mu - \alpha_{i,j,k} \geq 0, \quad \text{or} \tag{22}$$

$$\xi_{i,j,k} = y_{i,j,k} = 0. \tag{23}$$

From Eq. (22) we have that $\mu \geq \alpha_{i,j,k}$. Since χ^* corresponds to the smallest μ for which $g_0 (\xi, y) \leq \mu$, we can now obtain an interval $[l_1, l_2]$ that is guaranteed to include the optimal

value χ^* by finding the minimum and maximum values of $\alpha_{i,j,k}$:

$$l_1 = \min_{i,j,k} \alpha_{i,j,k}, \tag{24}$$

$$l_2 = \max_{i,j,k} \alpha_{i,j,k}. \tag{25}$$

With such interval, we can now apply Algorithm 1 to find the optimal value χ^* for the energy minimization problem. Once such value is found, the energy minimization problem can be expressed as a single convex feasibility problem:

$$find \quad \xi, y, \tag{26a}$$

$$subject\ to \quad \xi_{i,j,k} r_i \ln(2) - \beta \rho_{j,k} y_{i,j,k} [\chi^* - \alpha_{i,j,k}] \leq 0,$$

$$\forall i,j,k, \tag{26b}$$

$$\text{Constraints (8b)-(8h).} \tag{26c}$$

However, the optimum UE-to-CC association policy can be directly obtained from knowing χ^* without the need to solve this last optimization problem. From Eq. (22) and Eq. (23), we have that if $\chi^* \leq \alpha_{i,j,k}$, then $\xi_{i,j,k} = y_{i,j,k} = 0$. By plugging in the definition of $\alpha_{i,j,k}$ from Eq. (11), we can state in an equivalent way that the optimum UE-to-CC association policy for energy minimization is for UE i to associate with $\mathrm{CC}_{j,k}$ if and only if

$$\frac{h_{i,j,k}}{\eta_{i,j,k}} > w_{j,k} e^{-\chi^*}. \tag{27}$$

Since we obtained the above solution using the log-sum-exp approximation described in Eq. (13), it follows that the approximation gap is $\ln n$, where n is the product of the number of UEs, the number of layers, and the number of CCs per layer.

B. Load Balancing for Joint Energy Minimization and Capacity Maximization

In Section III-A, we analyzed the energy minimization problem in a HetNet where MSCA is supported. Even though the operators are able to reduce their economic and environmental impact by minimizing the energy consumption, they are typically interested in finding a balance between reducing the energy consumption and maximizing the network capacity. In this section, we analyze how these two conflicting objectives can be addressed jointly.

In a capacity maximization problem, the objective function generally follows the form of

$$f_1(\xi) = \sum_i r_i U \left(\sum_j \sum_k \xi_{i,j,k} \right), \quad \text{or} \tag{28}$$

$$f_2(\xi) = \sum_j U \left(\sum_k \sum_i r_i \xi_{i,j,k} \right), \tag{29}$$

where U is a concave function. A typical approach used in the literature is to consider U to be a logarithmic function. In such case, $f_1(\xi)$ represents a metric of fairness across multiple UEs, i.e., it is better to increase the bit rate of a user that is experiencing a low bit rate than to increase that of a user with an already high bit rate. Similarly, $f_2(\xi)$

represents a metric of load fairness across BSs, i.e., it is better to increase the total load (bit rate) carried by an underloaded BS than to increase the load of a BS that is already carrying a high load. Rather than focusing on a specific case, we will utilize a generic concave function $f_3(\xi)$. For such function, the capacity maximization problem can be expressed as

$$maximize \quad f_3(\xi), \tag{30a}$$

$$subject\ to \quad \sum_j \sum_k \xi_{i,j,k} \geq 1, \qquad \forall i, \tag{30b}$$

$$\sum_i y_{i,j,k} \leq 1, \qquad \forall j,k, \tag{30c}$$

$$\xi_{i,j,k} \geq 0, \qquad \forall i,j,k, \tag{30d}$$

$$y_{i,j,k} \geq 0, \qquad \forall i,j,k, \tag{30e}$$

$$r_i \xi_{i,j,k} - \rho_{j,k} \theta_{\max} y_{i,j,k} \leq 0, \quad \forall i,j,k. \tag{30f}$$

$$\xi_{i,j,k} = 0; \qquad \forall i,j,k, \mathrm{CC}_{j,k} \notin \Gamma_i \tag{30g}$$

$$y_{i,j,k} = 0; \qquad \forall i,j,k, \mathrm{CC}_{j,k} \notin \Gamma_i \tag{30h}$$

The only difference between the above constraints and the ones of the energy minimization problem is that here, the UE total QoS requirement r_i no longer needs to be satisfied with equality; rather, it is the lower bound, specified by constraint (30b). Therefore, the domain of the energy minimization problem is a subset of the one of the capacity maximization problem. Moreover, any feasible point for the energy minimization problem (8) is also feasible for the capacity maximization problem (30). Also, note that the inputs required from the UEs to solve the capacity maximization problem (30) are the same as for the energy minimization problem (8).

We can reformulate the capacity maximization problem as a convex minimization problem:

$$minimize \quad f_4(\xi) \equiv -f_3(\xi), \tag{31a}$$

$$subject\ to \quad \text{Constraints (30b)-(30h),} \tag{31b}$$

where $f_4(\xi)$ represents the new objective function. Since f_4 is the negative of a concave function, it is convex. If we denote by $f_0(\xi, y)$ the objective function of the energy minimization problem, then the problem of jointly minimizing the energy consumption and maximizing the network capacity can be expressed as

$$minimize \quad \begin{bmatrix} f_0(\xi, y) \\ f_4(\xi) \end{bmatrix}, \tag{32a}$$

$$subject\ to \quad \text{Constraints (30b)-(30h),} \tag{32b}$$

i.e., as a *multi-criterion* or *multi-objective* optimization problem. It is important to note that f_0 and f_4 are competing functions, i.e., one of them is minimized at the expense of increasing the other. Because of this competing nature, no single point is capable of jointly achieving the minimum value that f_0 and f_4 could achieve separately. However, since a multi-objective optimization is a vector optimization defined over a cone $K = \mathbb{R}_+^m$ for some $m > 0$, we can scalarize the problem to find Pareto-optimal points for the original problem.

Applying scalarization to the optimization problem (32), we obtain

$$\text{minimize} \quad v f_0\left(\xi, y\right) + (1-v)f_4\left(\xi\right), \tag{33a}$$
$$\text{subject to} \quad \text{Constraints (30b)-(30h),} \tag{33b}$$

where $0 \leq v \leq 1$ is a parameter that is adjusted to find the Pareto-optimal points. Intuitively, v is selected to indicate the operator's balance point between the energy minimization and the capacity maximization. For v close to 1, a greater weight is given to the energy minimization. Conversely, for v close to 0, a greater weight is given to the capacity maximization.

In general, for a given v, if f_0 and f_4 are convex functions, then the scalarized optimization problem is a convex one. We have shown that f_4 is a convex function, and, in Section III-A, we found that the energy minimization problem can be expressed as the convex feasibility problem (26). The issue with using the objective function of the latter is that, by definition, the objective function of a feasibility problem is a constant independent of the optimization variables. If we were to consider f_0 a constant, then it would have no effect on the solution of the optimization problem (33), i.e., such optimization problem would be reduced to the capacity maximization problem. Therefore, f_0 cannot be directly taken from the convex feasibility problem (26). However, we can obtain an appropriate f_0 from the original formulation of energy minimization problem, as we will now describe.

If we apply the weight factor v to the objective function of the original formulation of the energy minimization problem in (8), such problem becomes

$$\text{minimize} \quad v \sum_j \sum_k \hat{P}_{\text{on,dyn}}(\text{CC}_{j,k}), \tag{34a}$$
$$\text{subject to} \quad \text{Constraints (8b)-(8h),} \tag{34b}$$

which, after similar transformations as the one followed during the analysis of the energy minimization problem, becomes equivalent to

$$\text{minimize} \quad \ln\left(\sum_{i,j,k} \exp\left(\ln\left(v\right) + \frac{\xi_{i,j,k} r_i \ln\left(2\right)}{\beta \rho_{j,k} y_{i,j,k}} + \alpha_{i,j,k} \right) \right),$$
$$\tag{35a}$$
$$\text{subject to} \quad \text{Constraints (8b)-(8h).} \tag{35b}$$

Applying the log-sum-exp approximation described in Section III-A, the above optimization problem can be approximated as

$$\text{minimize} \quad \max_{i,j,k}\left[\ln\left(v\right) + \alpha_{i,j,k} + \frac{1}{\beta} \frac{\xi_{i,j,k} r_i}{\rho_{j,k} y_{i,j,k}} \ln\left(2\right) \right],$$
$$\tag{36a}$$
$$\text{subject to} \quad \text{Constraints (8b)-(8h).} \tag{36b}$$

For $v = 1$, the above problem is reduced to the original energy minimization problem. Therefore, for $v = 1$ and following a similar development as in Section III-A, the optimization

problem (36) is equivalent to a single convex feasibility problem

$$\text{find} \quad \xi, y, \tag{37a}$$
$$\text{subject to} \quad \xi_{i,j,k} r_i \ln\left(2\right) - \beta \rho_{j,k} y_{i,j,k}\left[\left(\chi^* - \ln\left(v\right)\right)\right.$$
$$\left. -\alpha_{i,j,k}\right] \leq 0, \qquad \forall i,j,k, \tag{37b}$$
$$\text{Constraints (8b)-(8h).} \tag{37c}$$

If we consider $0 < v < 1$ in the above problem, we note that its impact translates into increasing the effective threshold $\left(\chi^* - \ln\left(v\right)\right)$ of the optimization problem, since $\ln\left(v\right) < 0$. More importantly, in the above feasibility problem, the factor v is part of the constraint rather than of the objective function. Therefore, the problem above does not suffer from v not impacting the optimization problem (33), as was the case when we directly used problem (30). So, combining the objective function and constraints of the above problem with that of the capacity maximization as per the formulation of problem (33), we obtain that the scalarized multi-objective optimization becomes

$$\text{minimize} \quad -f_3\left(\xi\right), \tag{38a}$$
$$\text{subject to} \quad \xi_{i,j,k} r_i \ln\left(2\right) - \beta \rho_{j,k} y_{i,j,k}\left[\left(\chi^* - \ln\left(v\right)\right)\right.$$
$$\left. -\alpha_{i,j,k}\right] \leq 0, \qquad \forall i,j,k, \tag{38b}$$
$$\text{Constraints (30b)-(30h).} \tag{38c}$$

From the above problem formulation, we can also directly obtain the optimum UE-to-CC association policy without the need to find the solution. As in the case of the energy minimization problem, there are two possible necessary conditions for constraint (38b) to be satisfied:

$$\left(\chi^* - \ln\left(v\right)\right) - \alpha_{i,j,k} > 0, \quad \text{or} \tag{39}$$
$$\xi_{i,j,k} = y_{i,j,k} = 0. \tag{40}$$

Therefore, if $\left(\chi^* - \ln\left(v\right)\right) \leq \alpha_{i,j,k}$, then $\xi_{i,j,k} = y_{i,j,k} = 0$. By plugging in the definition of $\alpha_{i,j,k}$ from Eq. (11), we can state in an equivalent way that the optimum UE-to-CC association policy for any given v in the multi-objective optimization of energy minimization and capacity maximization is for UE i to associate with $\text{CC}_{j,k}$ if and only if

$$\frac{h_{i,j,k}}{\eta_{i,j,k}} > w_{j,k} \exp\left(-\left(\chi^* - \ln\left(v\right)\right) \right). \tag{41}$$

Once the UEs associate with the CCs, the values of the optimization variables ξ and y will depend on the particular function f_3 utilized as objective function of the capacity maximization problem. It is important to note that, to perform the multi-objective optimization, we need to find the value of χ^* only once, and then the UE-to-CC association policy is defined by that value, the operator-defined v, and the specific capacity maximization function of interest.

IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of our proposed algorithms for MSCA-enabled HetNets to minimize the energy consumption and balance it with the capacity maximization. The simulation parameters are shown in Table I. We have

TABLE I
SIMULATION PARAMETERS FOR MULTI-LAYER HETNETS WITH MSCA.

Parameter	Value
CC per BS (per layer)	[1,1,1] CC
Bandwidth of a CC (per layer)	[20, 10, 2.5] MHz
Number of antennas at BS	1
Max. spectral efficiency (θ_{max})	4.8 bps/Hz
Total coverage area	1km x 1km
β	0.75
Number of active UEs	50
Altitude of UEs	1.5 m
Number of antennas at UE	1
Number of layers	3
Type of BSs (per layer)	[macro,pico,pico]
Number of BSs (per layer)	[1,5,15]
Altitude of BSs (per layer)	[25,20,10]m
Power weight of a CC (per layer)	[2.66,3.1,4.0]

Fig. 2. Percent of UEs using MSCA and mean UE spectral efficiency.

chosen to assign only one CC to every BS in every layer so that our results capture the effects of MSCA rather than those of the classical CA. Based on the bandwidth parameters, we have that layer 1 (L1), layer 2 (L2), and layer 3 (L3) provide 18.6%, 46.51%, and 34.88% of the network capacity, respectively. Thus, L1 is meant to provide basic coverage, L2 is meant to provide basic capacity, and L3 is meant to enhance the capacity. BSs per layer and active UEs are uniformly distributed across the total coverage area.

For the path loss, we use the following 3GPP models for heterogeneous networks in outdoor scenarios (distance D (in km)) [47]:

$$(\text{macro})PL = 128.1 + 37.6\log(D), \qquad (42)$$

$$(\text{pico})PL = 140.7 + 36.7\log(D). \qquad (43)$$

To evaluate the overall performance of the energy-saving algorithm, we applied it to 100 different scenarios generated using the parameters from Table I. For each scenario, Algorithm 1 was evaluated for a minimum QoS varying in the range $[1, 10]$Mbps. To solve the convex feasibility problem in step 4 of Algorithm 1, we utilized CVX [48] [49], a Matlab-based modeling system for convex optimization, together with MOSEK [50], one of the leading commercial software products for large-scale optimization problems. In 91% of the 1000 scenario-QoS combinations and a lower bound $\epsilon = 0.01$, our algorithm required 11 iterations to converge to a solution and 12 iterations in the rest of the cases. This result highlights the high convergence rate achieved by the initial estimation of the interval $[l_1, l_2]$ from Eq. (24) and Eq. (25).

Figure 2 depicts the percent of UEs using MSCA and the mean UE spectral efficiency. From this figure, we observe that the percent of UEs using MSCA increases from less than 5% to nearly 35% as the minimum QoS requirement increases from 1 to 10Mbps. For this same variation of the minimum QoS requirement, the mean UE spectral efficiency grows from nearly 2.5bps/Hz to 4.7bps/Hz, i.e., it almost reaches the maximum spectral efficiency of 4.8bps/Hz.

The fact that most UEs are operating at nearly maximum spectral efficiency prevents more UEs from applying MSCA since such event would require a first set of UEs to empty part of its currently allocated spectrum so that a second set of
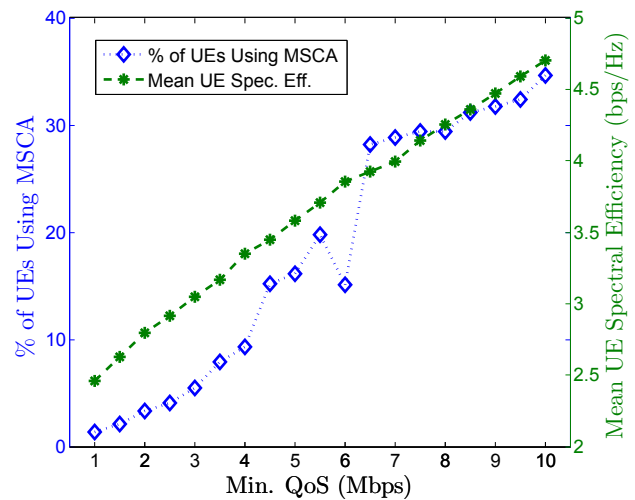
UEs, currently connected to other layers, can utilize the freed spectrum. However, such release of spectrum would imply that the UEs of the first set have to further increase their own spectral efficiency.

Figure 3 shows several per-layer metrics. In Figure 3a, we depict how the UEs associate with each BS layer. Here, we observe that as the minimum QoS requirement increases, the percent of UEs associated with L1 and L3 experiences small variations, indicating that most UEs remain connected to those layers. However, the percent of UEs attached to L2, the layer with highest capacity, increases significantly, indicating that most UEs are applying MSCA by connecting to an additional CC in L2. In Figure 3b, we observe that as the minimum QoS requirement increases, the percent of the load carried by L3 decreases from 51% to 34% while that of L2 increases from 24% to 47%. From Figure 3c, we observe that the change in the load managed by L2 and L3 produces a nearly equivalent change in the percent of energy consumption. That of L3 decreases from 30% to 19% while that of L2 increases from 35% to 56%.

An additional metric of interest is the value of χ^* as the minimum QoS requirement increases. This behavior is depicted in Figure 4. When the minimum QoS requirement is less than 7Mbps, χ^* increases almost linearly from -1.4 to 0.05. However, beyond 7Mbps, χ^* increases rapidly until reaching a value of 2.44.

To quantify the amount of energy savings provided by our energy-saving algorithm, as well as to characterize the energy-capacity trade-off in an MSCA-enabled HetNet, we take a single instance of a HetNet generated with the parameters of Table I, and analyze its performance as the factor υ varies from 0 to 1, representing a shift from the capacity maximization to the energy minimization objective. We analyze the balance of energy minimization against three objective functions for the capacity maximization: classical capacity maximization ($f_{3,1}$), global UE fairness ($f_{3,2}$), and per-BS UE fairness ($f_{3,3}$). Their respective definitions are as follows:
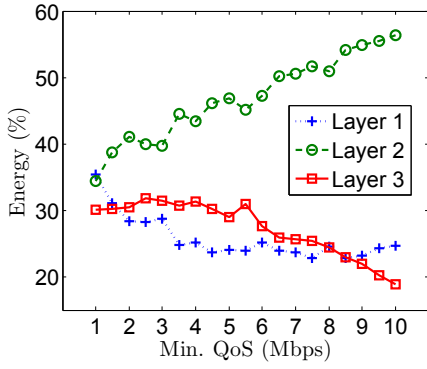
$$f_{3,1}(\xi) = \sum_i \sum_j \sum_k r_i \xi_{i,j,k}, \qquad (44)$$
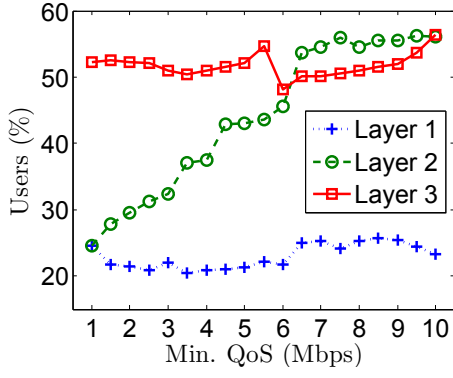
(a) Users per layer.
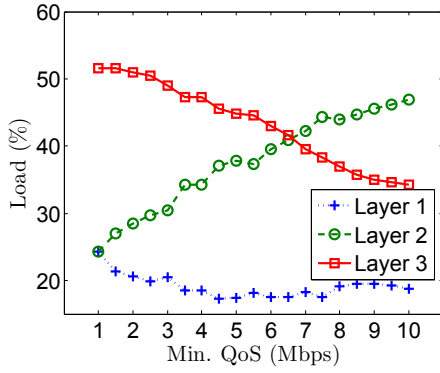
(b) Load per layer.

(c) Energy per layer.

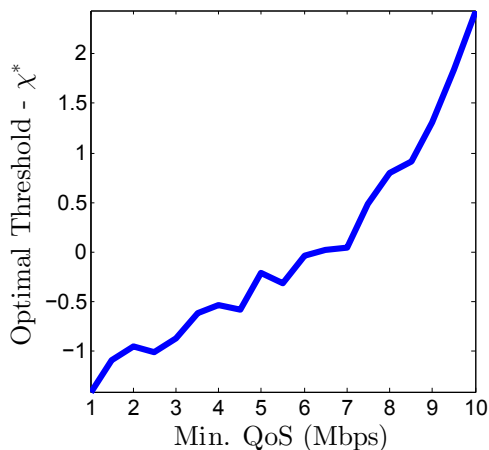Fig. 3. Users, load, and energy per layer.

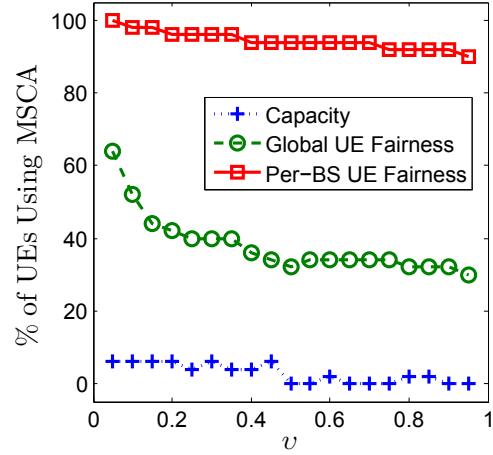Fig. 4. χ^* vs. minimum QoS requirement.

Fig. 5. MSCA UEs in the energy-capacity optimization.

$$f_{3,2}(\xi) = \sum_i \log \left(\sum_j \sum_k r_i \xi_{i,j,k} \right), \qquad (45)$$

$$f_{3,3}(\xi) = \sum_j \sum_i \log \left(\sum_k r_i \xi_{i,j,k} \right). \qquad (46)$$

In Figure 5, we show the percent of UEs that use MSCA. For the classical capacity maximization objective, we observe that very few UEs apply MSCA regardless of the value of υ. This behavior occurs because such objective tends to favor UE-to-CC links with higher SINR; therefore, MSCA links with distant BSs tend to be disregarded. On the other extreme, we have the per-BS UE fairness. In this case, the number of UEs applying MSCA is over 90%. This behavior occurs because each BS tries to provide a fair amount of throughput to all the UEs that it can potentially serve; therefore, this objective function encourages the application of MSCA among all the UEs that are under the coverage of more than one layer. We observe that the global UE fairness, with the use of MSCA decreasing from 64% to 30% as υ varies from 0 to 1, falls roughly in the middle between the other two extremes - capacity and per-BS UE fairness. These three graphs suggest that reconfiguring the balance between energy minimization and capacity maximization, i.e., changing the value of υ, will have the greatest impact on the UE-to-CC association when the objective function is that of Global UE Fairness. Thus, the network stability would require more attention in such reconfiguration scenario than when the other objective functions are used.

In Figure 6, we depict the capacity usage and the energy consumption. From Figure 6a, we observe that by applying the energy minimization algorithm it is possible to decrease the energy consumption to at least 15% of its maximum for all the capacity objectives. The effect of υ on the network capacity usage is shown in Figure 6b. From this graph, we observe that minimizing the energy consumption has the greatest impact, from 90% to 68%, on the capacity usage for the per-BS UE fairness objective. Conversely, the classical capacity objective experiences the least impact, from 99% to 91.45%. In Figures 6a and 6b, it is important to notice that the greatest
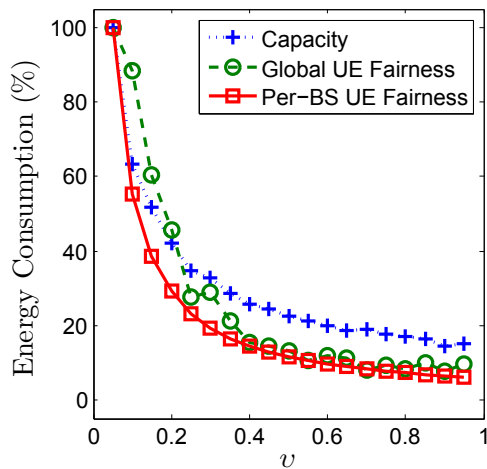
(a) Energy consumption.
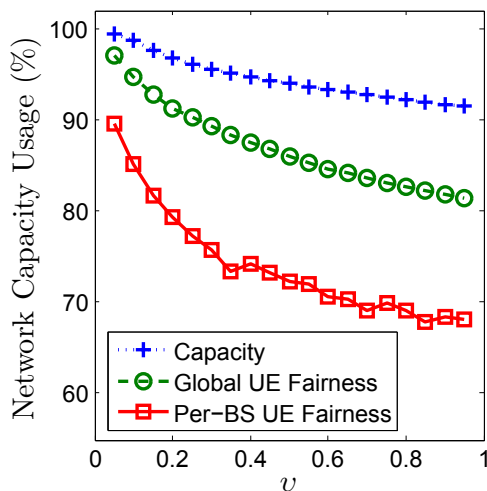
(b) Network capacity usage.

Fig. 6. Energy consumption and capacity usage in the energy-capacity optimization.

changes in energy consumption and energy capacity usage occur as the value of υ increases from zero to approximately 0.4; nonetheless, in such interval the rate at which the energy consumption decreases is faster than that at which the capacity usage decreases. This suggests that a good trade-off between both objectives can be achieved.

From the above graphs, we can now generate the energy-capacity trade-off curve for the MSCA-enabled HetNet, as shown in Figure 7. From this graph, we observe that indeed a good trade-off is achievable between both objectives: reducing the capacity usage by as little as 5% allows to significantly increase the energy savings in all of the three capacity objectives. However, even though it is possible to augment the energy savings by further reducing the capacity usage, the return from such reduction tends to diminish, particularly for the per-BS UE fairness objective.

V. CONCLUSIONS

MSCA has been introduced as a mechanism to increase the amount of bandwidth available to the users for HetNets in 5G cellular systems. However, existing work has focused on
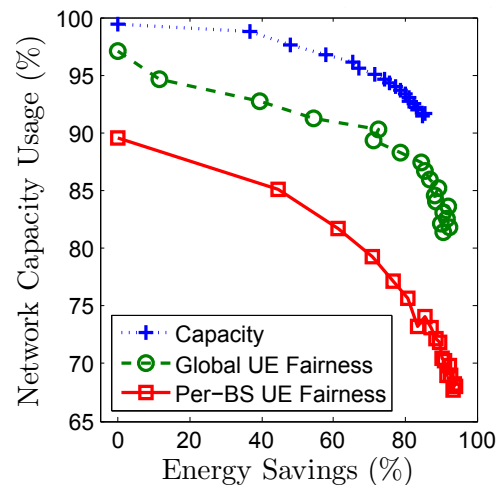

Fig. 7. Trade-off curve for the energy savings vs. capacity usage.

exploiting the use of MSCA to maximize the network capacity, disregarding the energy efficiency of MSCA. In this paper, we studied the problem of minimizing the energy consumption in MSCA-enabled HetNets and developed an efficient algorithm to solve it. We showed that, by utilizing a quasiconvex relaxation, we are able to not only solve the problem, but also to establish a clear and simple cell-association policy. Moreover, we showed how this cell-association policy can be easily adjusted to obtain a new policy that balances the conflicting objectives of energy minimization and capacity maximization. Through extensive simulations, we characterized the effects of our algorithm on the percent of load, users, and energy per layer, as well as on the percent of UEs that use MSCA and their average spectral efficiency. In addition, we obtained the trade-off curve between the energy minimization and capacity maximization and found that a large amount of energy savings can be achieved in an MSCA-enabled HetNet by reducing the network capacity usage by as little as 5%.

ACKNOWLEDGMENT

REFERENCES

[1] International Telecommunication Union, "Requirements related to technical performance for IMT-Advanced radio interface(s)," Report ITU-R M.2134, 2008.

[2] I. F. Akyildiz, D. M. Gutierrez-Estevez, R. Balakrishnan, and E. Chavarria-Reyes, "LTE-Advanced and the evolution to Beyond 4G (B4G) systems," *Physical Communication (Elsevier) Journal*, vol. 10, no. 0, pp. 31 – 60, 2014.

[3] *Radio Regulations*, International Telecommunication Union Std., Nov. 2012. [Online]. Available: http://www.itu.int/pub/R-REG-RR/en

[4] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2602336, IEEE Transactions on Wireless Communications

11

[5] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[6] K. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. Garcia, and Y. Wang, "Carrier aggregation for LTE-advanced: functionality and performance aspects," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 89–95, Jun. 2011.

[7] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger, and F. Xu, "Overview of 3GPP LTE-advanced carrier aggregation for 4G wireless communications," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 122–130, Feb. 2012.

[8] 3GPP, "Service accessibility," Technical Specification 22.011, Dec. 2011.

[9] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfro{}jd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5g," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 44–51, May 2014.

[10] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5g of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.

[11] A. Toskala, "Release 12 for c4 (cost, coverage, coordination with small cells and capacity)," in *TSG Ran Workshop on Rel-12 and onwards*. 3GPP, Jun. 2012.

[12] Huawei Technologies, "Views on Rel-12 and onwards for LTE and UMTS," in *TSG Ran Workshop on Rel-12*. 3GPP, Jun. 2012.

[13] NTT Docomo, "Requirements, Candidate Solution & Technology Roadmap for LTE Rel-12 Onward," in *TSG Ran Workshop on Rel-12*. 3GPP, Jun. 2012.

[14] C. Hoymann, D. Larsson, H. Koorapaty, and J.-F. Cheng, "A Lean Carrier for LTE," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 74–80, Feb. 2013.

[15] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, Sep. 2011.

[16] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "alpha-optimal user association and cell load balancing in wireless networks," in *Proc. of the IEEE International Conference on Computer Communication (INFOCOM)*, Mar. 2010, pp. 1–5.

[17] Y. Wang, K. Pedersen, T. Sorensen, and P. Mogensen, "Carrier load balancing and packet scheduling for multi-carrier systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1780–1789, May 2010.

[18] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[19] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-Efficient Scheduling and Power Allocation in Downlink OFDMA Networks with Base Station Coordination," *IEEE Trans. Wireless Commun.*, to be published, early Access.

[20] R. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[21] M. Ismail and W. Zhuang, "Network cooperation for energy saving in green radio communications," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 76–81, Oct. 2011.

[22] O. Holland, A. Aghvami, T. Dodgson, and H. Bogucka, "Intra-operator dynamic spectrum management for energy efficiency," *IEEE Commun. Mag.*, vol. 50, no. 9, pp. 178–184, Sep. 2012.

[23] B. Soret, H. Wang, K. Pedersen, and C. Rosa, "Multicell cooperation for LTE-advanced heterogeneous network scenarios," *IEEE Wireless Commun. Mag.*, vol. 20, no. 1, pp. 27–34, Feb. 2013.

[24] C. D. T. Thai, P. Popovski, M. Kaneko, and E. de Carvalho, "Multi-Flow Scheduling for Coordinated Direct and Relayed Users in Cellular Systems," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 669–678, Feb. 2013.

[25] M. Simsek, M. Bennis, and I. Guvenc, "Enhanced intercell interference coordination in hetnets: Single vs. multiflow approach," in *IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 725–729.

[26] ——, "Learning Based Frequency- and Time-Domain Inter-Cell Interference Coordination in HetNets," *IEEE Trans. Veh. Technol.*, to be published, early Access.

[27] X. Lin, J. Andrews, R. Ratasuk, B. Mondal, and A. Ghosh, "Carrier aggregation in heterogeneous cellular networks," in *IEEE International Conference on Communications (ICC)*, Jun. 2013, pp. 5199–5203.

[28] X. Lin, J. Andrews, and A. Ghosh, "Modeling, Analysis and Design for Carrier Aggregation in Heterogeneous Cellular Networks," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 4002–4015, Sep. 2013.

[29] 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 Commun. Mag.*, vol. 52, no. 2, pp. 122–130, 2014. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6736752

[30] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5g perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[31] W. Liu, S. Han, and C. Yang, "Hybrid cooperative transmission in heterogeneous networks," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2012, pp. 921–925.

[32] K. Huang, S. Lu, and J. Guo, "An Optimized Cooperative Transmission Scheme for Interference Mitigation in Heterogeneous Downlink Network," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, Sep. 2012, pp. 1–5.

[33] B. Clerckx, Y. Kim, H. Lee, J. Cho, and J. Lee, "Coordinated multi-point transmission in heterogeneous networks: A distributed antenna system approach," in *Proc. of the IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2011, pp. 1–4.

[34] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 44–50, Nov. 2012.

[35] Q. Li, R. Hu, Y. Qian, and G. Wu, "Cooperative communications for wireless networks: techniques and applications in LTE-advanced systems," *IEEE Wireless Commun. Mag.*, vol. 19, no. 2, Apr. 2012.

[36] P. Marsch and G. Fettweis, "A decentralized optimization approach to backhaul-constrained distributed antenna systems," in *Proc. of the IST Mobile and Wireless Communications Summit (ISTMWC)*, Jul. 2007, pp. 1–5.

[37] ——, "A framework for optimizing the uplink performance of distributed antenna systems under a constrained backhaul," in *Proc. of the IEEE International Conference on Communications (ICC)*, Jun. 2007, pp. 975–979.

[38] 3GPP, "Enhanced CoMP for LTE," Samsung, Work Task RP-121803, Dec. 2012.

[39] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.

[40] F. Richter, A. Fehske, P. Marsch, and G. Fettweis, "Traffic Demand and Energy Efficiency in Heterogeneous Cellular Mobile Radio Networks," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, May 2010, pp. 1–6.

[41] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2010, pp. 1660–1664.

[42] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, C. Desset, and O. Blume, "Cellular Energy Efficiency Evaluation Framework," in *Proc. of the IEEE Vehicular Technology Conference (VTC)*, May 2011, pp. 1–6.

[43] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. Gonzalez, H. Klessig, I. Godor, M. Olsson, M. Imran, A. Ambrosy, and O. Blume, "Flexible power modeling of LTE base stations," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2012, pp. 2858–2862.

[44] E. Chavarria-Reyes and I. F. Akyildiz, "Radio Access Network Energy Minimization in Multi-Layer Heterogenoeus Wireless Systems," in *Proc. of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2013, pp. 3259–3263.

[45] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios," Technical Report 36.942, Sep. 2012.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[47] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," Technical Report 36.814, Mar. 2010.

[48] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[49] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2602336, IEEE Transactions on Wireless Communications

12

and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[50] ——, "Using mosek with cvx," http://web.cvxr.com/cvx/doc/mosek.html, Mar 2014.

Elias Chavarria Reyes received the B.E. degree in Electronics and Communication Engineering from Universidad de Panamá, Ciudad de Panamá, Panamá, in 2007. In May 2010, he received his M.S. degree in Electrical and Computer Engineering from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. He completed his PhD in 2014 with the Broadband Wireless Networking Lab at the Georgia Institute of Technology, under the supervision of Prof. Dr. Ian F. Akyildiz, with a fellowship of "SENACYT". His research focused on energy modeling and analysis in heterogeneous wireless systems. He is currently a software engineer with the Chief Technology Architecture Office at Cisco focusing on cable networks evolution. He is a member of IEEE.
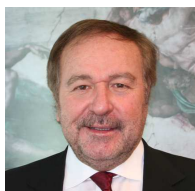
Ian F. Akyildiz received the B.S., M.S., and Ph.D. degrees in Computer Engineering from the University of Erlangen-Nrnberg, Germany, in 1978, 1981 and 1984, respectively. Currently, he is the Ken Byers Chair Professor in Telecommunications with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, the Director of the Broadband Wireless Networking Laboratory and Chair of the Telecommunication Group at Georgia Tech. Dr. Akyildiz is an honorary professor with the School of Electrical Engineering at Universitat Politècnica de Catalunya (UPC) in Barcelona, Catalunya, Spain and the founder of N3Cat (NaNoNetworking Center in Catalunya). Since 2011, he is a Consulting Chair Professor at the Department of Information Technology, King Abdulaziz University (KAU) in Jeddah, Saudi Arabia. Since January 2013, Dr. Akyildiz is also a FiDiPro Professor (Finland Distinguished Professor Program (FiDiPro) supported by the Academy of Finland) at Tampere University of Technology, Department of Communications Engineering, Finland. He is the Editor-in-Chief of Computer Networks (Elsevier) Journal, and the founding Editor-in-Chief of the Ad Hoc Networks (Elsevier) Journal, the Physical Communication (Elsevier) Journal and the Nano Communication Networks (Elsevier) Journal. He is an IEEE Fellow (1996) and an ACM Fellow (1997). He received numerous awards from IEEE and ACM. His current research interests are in nanonetworks, Long Term Evolution (LTE) advanced networks, cognitive radio networks and wireless sensor networks.

Etimad Fadel Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia. Received her Bachelor's degree in Computer Science at King Abdul Aziz University with Senior Project title ATARES: Arabic Character Analysis and Recognition in 1994. She was awarded the Mphil/PhD degree in computer science at De Montfort University (DMU), Liecester, UK. The Thesis title was Distributed Systems Management Service in 2007. Currently she is working as an Assistant Professor at the Computer Science Department at KAU. During this time she was appointed in different administrative positions. She was the Vice-Dean of the Faculty FCIT at the Girls section. This was between 2008 2010. Her main research interest is Distributed Systems, which are developed based on middleware technology. Currently she is looking into and working on Wireless Networks, Internet of Things and Internet of Nano-Things. In addition, she is working on smart grids and HetNets.