Spectrum-Energy Efficiency Optimization for Downlink LTE-A for Heterogeneous Networks

Chan-Ching Hsu and J. Morris Chang, *Senior Member, IEEE*

Abstract—Heterogeneous networks have been pointed out to be one of the key network architectures that help increase system capacity and reduce power consumption for efficient communications. Although conceivably, high operational efficiency brings a high profit for mobile service providers, it is noteworthy that the potential for maximizing the profit has not been explored for the heterogeneous environment. This paper investigates profitability for network operators with the spectrum-energy efficiency metric on the downlink of LTE Advanced communication systems. We pursue optimal policies by employing the techniques of cell size zooming, user migration and sleep mode in the deployment of different base station types. The problem is formulated as a quasiconvex optimization problem and it is transformed into an equivalent form of the MILP problem; the former is solved with a bisection algorithm and the latter is approached by an off-the-shelf software package. Since the formulated optimization problem is NP hard, a sub-optimal approach with a lower computational complexity is also proposed. Numerical analysis through case studies are presented to evaluate the efficiency improvements, and demonstrate the performance of the near-optimal solution.

Index Terms—Network optimization, energy-efficient, spectrum efficiency, heterogeneous network.

1 Introduction

TOTAL mobile traffics of the whole mobile world are growing fast; the increasing demand over the last mile access networks has motivated the network operators to extend and upgrade infrastructure in order to provide guaranteed coverage and data services. Fourth generation technologies such as LTE cellular systems have been developed and are expected to be important technologies to improve end-user throughput and network capacity. With such technological capability to improve information services, mobile operators have been experiencing annual increases in data traffic volumes and so in revenues. Meanwhile, [1] points out that the operational expenditure accounts for more than 18% of the energy bill, and that 60% of the power consumption in the network equipment is from base stations (BSs). The increasing data amount raises the profit at the expense of energy consumption increased.

From the mobile operator perspective, it is of interest to operate wireless networks in an economically efficient fashion. It was analyzed in [2] that deploying different types of BSs is able to help on expenditure reduction and increase revenues. Thus, a hybrid of cellular deployments provides great leeway to the network operators to attain financial improvements [3]. Nevertheless, one of the concerns with HetNets is the growing number of small cell sites. The total energy consumption of all femtocells will amount more than 3.784×10^9 kWh/annum according to [4]. This raises important questions about the energy efficiency implications of HetNet deployment. Hence, energy efficient operation in a heterogeneous setting becomes a pressing issue in order to reduce the energy expenditure.

On the other hand, when evaluating the energy saving mechanisms, the impact on spectrum efficiency should also be taken into account. The authors in [5] analyzed the tradeoff between the two metrics. The derived analytical expression shows that the increase in energy efficiency will inevitably bring down the spectrum efficiency. Schemes aimed at joint optimization of energy efficiency and spectral efficiency should be well studied. However, these two network efficiency improvement problems are focused and studied disjointedly in the literature. Therefore, in this paper, a spectrum-energy efficiency optimization model is developed for the heterogeneous network environments. We define maximizing network spectrum-energy efficiency as optimizing the spectrum allocation while minimizing the total energy consumption.

There is a long-term economic objective of spectrum-energy efficiency optimization. The spectrum-energy efficiency can be considered as a metric to quantify the revenue-per-cost capacity, serving as an indicator of operators profitability. It is a generalized objective function that maximizes the ratio of spectrum efficiency over energy consumed for the service operation. The goal is to attain a network design which is cost-effective, improving the data rate per resources managed (resource blocks and energy). As an efficient network is capable of growing in data volume conveyed successfully, within given cost in terms of energy consumption, the revenue then increases, coming from expanded data services offered. With the optimal service rate per resource, the most profitable network design can be realized.

In this research, the spectrum-energy efficiency in

• *Chan-Ching Hsu and J. Morris Chang are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa, 50011.*
E-mail: cchsu@iastate.edu, morris@iastate.edu

cellular communication systems is defined as the bits/s-per-RB-per-Joule capacity, which depends on the served data rate, allocated transmit power, resource blocks (RBs) and static BS energy consumption. We propose to migrate user equipment (UE) (i.e., mobile devices), switch OFF and zoom in/out BSs. To obtain the optimal operation, a quasiconvex programming problem is developed, and is transformed into a Mixed Integer Linear Programming problem (MILP), where considered are association, spatial, resource, operation and service constraints. The decision space contains BS-UE-Modulation association assignments and BS operation mode.

The contributions are summarized as follows: (i) We present an analytical framework for optimal BS operation, user association and modulation scheme selection in HetNets. Our objective is to maximize spectrum-energy efficiency, i.e., maximizing profitability of network service providers. We stress on techniques to enhance the utilization of resource blocks and thus to better amortize the license costs of frequency bands for HetNets. (ii) Two heuristic approaches are proposed to find near-optimal solutions. The bisection method produces a solution different from the optimal within a predefined tolerable value; the devised heuristic algorithm reaches a sub-optimal solution with computational tractability. We run extensive simulations based on non-uniform BS topologies and traffic distribution scenarios to verify performance of the optimization framework, and heuristics.

The remainder of this paper is organized as follows. We review related work in the next section; in Section 3, a system model is presented which portrays the studied heterogeneous network environment. The spectrum-energy efficiency optimization scheme is then proposed in Section 4. The heuristic approaches in Section 5 are proposed to enhance the computation efficiency, followed by the numerical experiments in Section 6. Finally, Section 7 concludes this paper.

TABLE 1
List of Key Mathematical Notations

$\mathbb{B}_{\mathbb{A}}$	set of macro cells, a_m
$\mathbb{B}_{\mathbb{O}}$	set of femto cells, o_f
$\mathbb{B}_{\mathbb{S}}$	set of BSs, $\{b_s\} = \{a_1, \ldots, a_M, o_1, \ldots, o_F\}$
$\mathbb{Q}_{\mathbb{K}}$	set of modulation techniques, q_k
$\mathbb{U}_{\mathbb{L}}$	set of users, $\{u_l\} = \{x_1, \ldots, x_N, w_1, \ldots, w_E\}$
$\mathbb{C}$	BS-UE connectivity matrix, $c_l^s = \{0, 1\}, \forall s \in \mathbb{B}_{\mathbb{S}}, l \in \mathbb{U}_{\mathbb{L}}$
$\mathbb{I}$	traffic demand matrix, $i_l, \forall l \in \mathbb{U}_{\mathbb{L}}$
$\mathbb{P}$	transmit power matrix, $p_{lk}^s, \forall s \in \mathbb{B}_{\mathbb{S}}, l \in \mathbb{U}_{\mathbb{L}}, k \in \mathbb{Q}_{\mathbb{K}}$
$\mathbb{D}$	resource block number matrix, $r_{lk}, \forall l \in \mathbb{U}_{\mathbb{L}}, k \in \mathbb{Q}_{\mathbb{K}}$
$\mathbb{V}$	set of transmit power parameters, $v_l^s, \forall s \in \mathbb{B}_{\mathbb{S}}, l \in \mathbb{U}_{\mathbb{L}}$
d_{sl}	distance between BS s and user l
L	total number of users, $\lvert\mathbb{U}_{\mathbb{X}}\rvert + \lvert\mathbb{U}_{\mathbb{W}}\rvert$
R_b	network blocking ratio
S	total number of BSs, $\lvert\mathbb{B}_{\mathbb{A}}\rvert + \lvert\mathbb{B}_{\mathbb{O}}\rvert$
$\mathbf{H_m}$	set of femto cells in the macro cell, $h_g^m = f, \forall m \in \mathbb{B}_{\mathbb{A}}$
$\mathbf{P_{active}}$	basic active energy consumption vector, $p_{active}^s, \forall s$
$\mathbf{P_{sleep}}$	sleep mode energy consumption vector, $p_{sleep}^s, \forall s \in \mathbb{B}_{\mathbb{S}}$
$\mathbf{P_{max}}$	maximum total transmit power vector, $p_{max}^s, \forall s \in \mathbb{B}_{\mathbb{S}}$
$\mathbf{R}$	maximum number of RB vector, $R^m, \forall m \in \mathbb{B}_{\mathbb{A}}$
$\mathcal{Y}$	BS-UE-Modulation association matrix, $(y_{lk}^s)_{S \times L \times K}$
$\mathcal{Z}$	BS operation mode vector, $(z_s)_{1 \times S}$

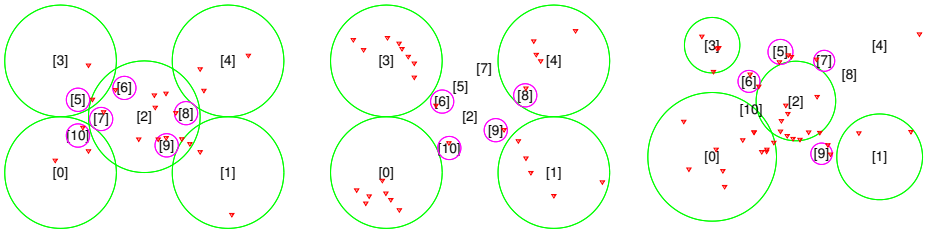
2 Related Work

Several approaches have been pursued to reduce BSs' energy consumption. Dynamic BS operation (i.e., switching on/off scheme) has been investigated, allowing a significant amount of energy to be saved, motivated by the traffic load fluctuation [6]–[9]. Large savings, depending on temporal-spatial traffic dynamics, are shown to be possible. [10] finds the minimal transmission power that ensures coverage and capacity, but is not sufficient to reduce the energy consumption due to the load-independent components of the energy consumption. [11] proposes centralized and distributed algorithms to verify the reduction of power consumption of a cellular network with cell zooming, and to avoid coverage hole when BSs are turned off. [12] does not consider power for active/sleep mode and switches between two radii that minimize consumption depending on the traffic at every hour, which is a simple case. [13] compares different cell sizes and concludes that smaller cells are more favorable for high energy efficiency; only transmit power, however, is counted therein. In our work, cell size does not alternate between fixed radii, and no predefined traffic threshold or fixed traffic volume is based on in our design so that responding to dynamics more effectively. We consider coordinations among stations with the optimization-based scheme.

Beside energy-efficient cell-size design, emerging heterogeneous networks have also received considerable attention in the recent years. A well-deployed heterogeneous network will not only bring better performance on coverage and capacity but also higher energy-efficiency [14]–[16]. However, network planning is primarily an off-line activity and is not necessarily fine to handle dynamic scenarios. Cloud radio access network (C-RAN) is also proposed, utilizing centralized processing, separated from the radio access units and dynamical network switching to address the efficiency issues [17]. However, a C-RAN with large-scale centralization may incur enormous fronthaul expenditure. In contrast, our approach is proposed as an enhancement to the existing LTE RAN. Our prior work [18] studies the spectrum-energy efficiency in only the homogeneous setting. In this work, beside the heterogeneous environment, the differences can be mentioned by the following points (i) considering the effects of modulation scheme, association and spatial constraints to formulate a more sophisticated model, and (ii) further proposing algorithms for the dynamic BS operation and user association as well as presenting analysis to justify the effectiveness.

3 Network Models

In our work, the broadband wireless communication system is considered which consists of different kinds of BSs



(a) Cells running conventionally (b) Cells switching to sleep mode (c) Cells varying the service area

Fig. 1. BS operations, cell zooming in/out, sleep mode and user migration, to enhance network performance

and manages multiple users. Spectrum-energy efficiency is used to assess network capacity and defined as served traffic over allocated resource block per energy consumed. The goals are to (i) maximize spectrum-energy efficiency, (ii) obtain the optimal BS operation mode, and (iii) assign serving BSs to users with designated modulation techniques that satisfy rate requirements of associated users. A list of the key mathematical notations is shown in Table 1.

Cell zooming and sleep mode operations as shown in Fig. 1 are of benefit to enhancing spectrum-energy efficiency. Fig. 1(a) shows a wireless access network formed with five macro (green circles) plus six femto cells (pink circles), where the central macro cell is surrounded by four macros and works with 4 femto cells.[1] BSs are located at the respective center of the cells, designated with numbers; users are distributed among BSs, denoted by red points. In Fig. 1(b) macro BS-2 and other femto cells work in sleep mode, together reducing energy consumption. An example is provided from Fig. 1(c) to illustrate cell zooming and user migration. Macro BS-0 and femto BS-5, BS-6 zoom out to associate with the users migrating from macro BS-2 or BS-3; consequently, macro BS-2 and BS-3 are allowed to reduce the coverage with decreased power. With the network blocking ratio, macro BS-4 chooses to sleep, disconnecting with the tolerable number of users. Moreover, connection through different modulation technique selections results in different levels of spectrum-energy efficiency.

In this work, when switched to sleep mode, BSs are not totally shut down and still have the ability to broadcast standard messages. BSs in sleep mode periodically broadcast beacon messages, and any user able to receive the messages can synchronize with BSs. When users are synchronized, necessary information is obtained and estimated. Based on the captured information, users are associated with BSs. In other words, the coverage shown in the figure is the BS association range instead of the broadcast range. Hence, blocked users are always allowed to request for service.

1. In our research, we consider femto-cells to represent small cells which are installed outdoor and managed by mobile operators. The result can be extended into HetNets consisting of other types of smaller sites such as micro BSs.

3.1 Overview

We present the optimization model for the BS spectrum-energy efficiency problem as follows. Let $\mathbb{U}_{\mathbb{L}} = \{u_1, \ldots, u_L\}$ be the set of UEs, and each UE device has a set of modulation schemes, $\mathbb{Q}_{\mathbb{K}}$, for data transmission. Let $\mathbb{B}_{\mathbb{A}} = \{a_1, \ldots, a_M\}$ be the set of macro base stations and $\mathbb{B}_{\mathbb{O}} = \{o_1, \ldots, o_F\}$ the set of femto base stations, respectively; $\mathbb{B}_{\mathbb{A}}$ is combined with $\mathbb{B}_{\mathbb{O}}$, introducing a new set comprising both types of base stations, $\mathbb{B}_{\mathbb{S}} = \{a_1, \ldots, a_M, o_1, \ldots, o_F\}$ with cardinality $|\mathbb{B}_{\mathbb{S}}| = S$.

Problem Statement: Given (i) transmission power, p^s_{lk}, between L users and a finite number of S stations with K modulation techniques, (ii) user rate requirements, i_l, required number of resource blocks, r_{lk}, (iii) blocking ratio, R_b, and (iv) maximum power, P^s_{max}, available resource blocks, R^m, basic active and sleep mode power usage of BSs, P^s_{active} and P^s_{sleep}, the problem is to maximize the system spectrum-energy efficiency, by associating users and BSs through designated modulation techniques. The decision space consisting of *association matrix* ($\mathcal{Y}$) and *operation mode vector* ($\mathcal{Z}$) is defined as follows:

- $\mathcal{Y} = \{y^s_{lk}\}$. Equals 1 if user l is assigned to base station s through modulation $k, \forall s \in \mathbb{B}_{\mathbb{S}}, l \in \mathbb{U}_{\mathbb{L}}, k \in \mathbb{Q}_{\mathbb{K}}$, and 0 otherwise.
- $\mathcal{Z} = \{z_s\}$. Equals 1 if base station s is in the active state, $\forall s \in \mathbb{B}_{\mathbb{S}}$, and 0 otherwise.

With information on power and carrier requirements, when y^s_{lk} is asserted, the traffic demand is satisfied for u_l via modulation technique k, and o_s has to be the value of one correspondingly. The BS cell size is the equivalent of the area within which BSs are able to serve users by tuning up transmission power (p^s_{lk}). Cell zooming is present by comparison of the service areas of two statues at consecutive operation times. In other words, as the service area varies with the adjusted transmit power level, the BS is undertaking cell zooming.

3.2 Association and Spatial Constraints

Eq. (1) realizes the traffic request from an associated user is satisfied through a direct link with a BS. Each user can be served by up to one BS through one modulation technique. For an individual BS s, if the result of $\sum_{l \in \mathbb{U}_{\mathbb{L}}, k \in \mathbb{Q}_{\mathbb{K}}} y^s_{lk}$ is equal to 0, it means s operates in sleep mode. Furthermore, the idea of user migration is put

into action if an user is associated with a BS which is different from the one it formerly connected.

$$\sum_{s\in\mathbb{B}_\mathbb{S},k\in\mathbb{Q}_\mathbb{K}} y^s_{lk} \leq 1, \forall l \in \mathbb{U}_\mathbb{L}, \tag{1}$$

Continuity Constraint: To prevent intermittent data transmission on the users, we distinguish existing users from new coming ones. The existing users are inclined to complete the ongoing sessions with continuous association sessions; the new coming users arrive at the network and seek for being associated with no data transmission yet. In this context, the set of users is partitioned into two, $\mathbb{U}_\mathbb{L} = \{\mathbb{U}_\mathbb{X}, \mathbb{U}_\mathbb{W}\}$ with cardinality $L = N + E$. The first part presents the set of new arrival users with $\mathbb{U}_\mathbb{X} = \{x_1, \ldots, x_N\}$; the second part represents the set of existing users with $\mathbb{U}_\mathbb{W} = \{w_1, \ldots, w_E\}$. To allow active users to be not service-discontinued, the following constraint is imposed. Note that when $\mathbb{U}_\mathbb{L} = \mathbb{U}_\mathbb{X}$, the association continuity is not taken into account.

$$\sum_{s\in\mathbb{B}_\mathbb{S},k\in\mathbb{Q}_\mathbb{K}} y^s_{ek} = 1, \forall e \in \mathbb{U}_\mathbb{W}, \tag{2}$$

Topology Constraint: The scheme of frequency reuse help BSs prevent interfering with the neighboring; as the cell size is adjustable, nevertheless, BSs could overexpand the transmission range, seriously interfering others sharing the same spectrum. Considering such impact, limiting individual transmit power is effective; therefore, whether user l is attainable for BS s can be determined in terms of the transmit power with modulation K, which requires the highest sensitivity and SNR among $\mathbb{Q}_\mathbb{K}$.

$$C(p^s_{lK}, p^s) \rightarrow c^s_l = \begin{cases} 1, \text{if } p^s_{lK} < p^s, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}; \\ 0, \text{otherwise}, \end{cases}$$

where p^s is a maximal individual transmit power, which can be estimated considering BSs were deployed at fixed locations. $c^s_l = 1$ means cell s can reach l in terms of transmit power. Then it is expressed as follows that BSs should only consider users within the reachable range as a feasible set.

$$\sum_{k\in\mathbb{Q}_\mathbb{K}} y^s_{lk} \leq c^s_l, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, \tag{3}$$

3.3 Power and Carrier Constraints

Two parameters are tied to modulation selection, transmit power and resource block. Given the traffic rate required at the user, the power and resource block number are obtained that are necessary for modulation schemes at the BS. Each modulation and coding scheme (MCS) is able to support a data rate; the transmission time is captured by the resource block number. Through modulations, a traffic requirement can be met by different numbers of resource blocks in given time duration, i.e. more resource blocks are allocated to achieve same data rate for lower MCS. We consider a quasi-static network scenario, where $i_l, \forall i \in \mathbb{U}_\mathbb{L}$, remains unchanged within every operation period, while changing across periods.

TABLE 2
Parameters per Modulation Scheme

Modulation	Sensitivity (dBm)	Data bit per symbol	Data rate (Mbps)
QPSK1/2	-96	1.0	2.2
16QAM1/2	-90	2.0	4.4
16QAM3/4	-87	3.0	6.6
64QAM3/4	-80	4.5	9.9

Individual rate requirements, i_l, are randomly chosen from the set of supported rates by the physical layer (Table 2). r_{lk} indicates the number of resource block for user l through modulation technique k.

Resource Block Constraint: The total number of resource blocks a BS serves should not be greater than the available number ($R^m, \forall m \in \mathbb{B}_\mathbb{A}$) on the channel bandwidth in a specified symbol time. This constraint is ensured by the following equation.

$$\sum_{h^m_g\in\mathbf{H_m},l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} [(y^m_{lk} + y^{h^m_g}_{lk}) \cdot r_{lk}] \leq R^m, \forall m, \tag{4}$$

As the available spectrum span is finite, the number of resource blocks to transmit data over the span is limited. We assume that non-overlapping spectrum bands are used, and hence there is no co-channel interference. In other words, orthogonal radio resources are allocated to macrocell and small cell users; the cross-tier and co-tier interference are completely eliminated. We further assume that small cells use orthogonal channels. Therefore, the different subsets of mobile users assigned to each small cell do not interfere with each other.

When operating within a macro cell's coverage, a femto cell allocates a portion of m's available frequency for transmissions. $\mathbf{H_m} = \{h^m_1, \ldots, h^m_{G_m}\}, \forall m \in \mathbb{B}_\mathbb{A}$ denotes the set of femtos f sharing spectrum with the macro m. To calculate the number of resource blocks for individual data transmissions with modulation technique k, the equation for the resource block number, r_{lk}, given the rate requirement, i_l, is considered as [19]:

$$R(i_l, k) \rightarrow r_{lk}, \forall l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K},$$

Transmission Power Constraints: Receiver sensitivity is regarded as the minimum power at the receiver side for the guaranteed signal strength, from which the necessary transmission power level, p^s_{lk}, can be calculated as according to [20]:

$$P(i_l, \mathbf{v^s_l}, k) \rightarrow p^s_{lk}, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K},$$

where $\mathbf{v^s_l}$ represents the propagation gains and attenuation by miscellaneous losses between s and l. It is shown the linear model can offer approximation for the transmission power with respect to the carried traffic load, and has been adopted in [12]. In this work, the model is applied to capture the dependency of traffic volume and resource blocks. BSs are considered to scale their power consumption to traffic load; a higher user

population implies higher transmit power from the base station generally. The constraints on total transmission power constraints are as follows:

$$\sum_{l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} (y^m_{lk}\cdot p^m_{lk}) \le P^m_{max}, \forall m \in \mathbb{B}_\mathbb{A}, \tag{5}$$

$$\sum_{l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} (y^f_{lk}\cdot p^f_{lk}) \le P^f_{max}, \forall f \in \mathbb{B}_\mathbb{O}, \tag{6}$$

Receiver sensitivity is listed in Table 2 [21]. As BS antennas are assumed to transmit ideally, feeder losses, connector losses and jumper losses are not considered. We consider fast fading margin and building penetration loss, whose values are assumed in simulation [22]. The COST-231 Hata and COST-Walfisch-Ikegami models are considered for path loss. We assumed the BS antenna height is 30(m) and the users' antenna height is 1(m).

We consider a scenario where for each cell, the input data for optimization are reported periodically, which is common in many real network deployments [23]. Besides, the timescales of changing parameters for all practical purposes are in the order of a few minutes (typically, 5-15 mins). Thus, reconfiguration ought to happen whenever: 1) input information changes significantly in some cells, or 2) when a maximum duration elapses since the last reconfiguration. In addition, if the estimated improvement upon new reconfiguration is small, then the previous configuration can be retained.

3.4 Operation and Service Constraints

If it is an BS operating in active mode, b_s is equal to one for s. Eq. (7) and (8) together ensure that a BS switches into active mode as long as delivering services; otherwise the BS switch to sleep mode without serving users.

$$z_s \le \sum_{l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} y^s_{lk}, \forall s \in \mathbb{B}_\mathbb{S}, \tag{7}$$

$$y^s_{lk} \le z_s, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K}, \tag{8}$$

The total number of connected users must be higher than the required number of active users as follows:

$$\sum_{s\in\mathbb{B}_\mathbb{S},l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} y^s_{lk} \ge \delta, \tag{9}$$

where $\delta = S(1-R_b)$ and R_b denotes the blocking ratio, the ratio of users not associated among total users demanding services. No single user is considered to be blocked if the rate is set to zero. Note that there could be no feasible solutions due to insufficient resources to serve users. In this case, operators can consider two alternatives: (i) increase the blocking ratio, (ii) treat existing users as new coming in the model. Having a high blocking ratio can lead to the situation where many new coming users have difficulty in getting mobile services; seeing existing users as new arrival can result in the situation where the service continuity is not guaranteed. On user side, they may experience a data rate drop overall, which commonly occurs when the network is congested. However, when available recourses are sufficient for data transmission, none of the alternatives is needed to be taken.

In LTE, BSs in sleep mode periodically wake up and broadcast beacon messages. The information regarding the power level adjustment is estimated and can be obtained through periodic channel quality and reference signal reports; the information about the number of RBs for user requirements can be collected during the resource allocation procedure. When users are able to synchronize with BSs, the information concerning transmit power can be used as input to our model without knowing the user locations. Hence, the values of p^s_{lk} and r_{lk} are parts of the parameters exchanges during communication establishment and maintenance. Once users have been associated, they will periodically send information to maintain communication.

4 Spectrum-energy Efficiency Problem

We define the spectrum efficiency as a summation of the individual served traffic divided by the allocated resource block number of associated users since the resources are distributed separately. The next expression presents the spectrum efficiency achieved:

$$\mathbf{S}(\mathcal{Y},\mathbb{I},\mathbb{D}) = \sum_{s\in\mathbb{B}_\mathbb{S},l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} (\frac{i_l}{r_{lk}}\cdot y^s_{lk}).$$

A BS operates in either sleep mode, taking a low energy level, or in active mode, consuming a basic running energy plus the energy for transmission. The energy a sleeping BS consume is P^s_{sleep} ; the energy consumed by an active BS comes from the basic active mode energy P^s_{active} plus the energy used to serve its users, $\sum_{l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}}(p^s_{lk}\cdot y^s_{lk}), \forall s \in \mathbb{B}_\mathbb{A}$. The overall energy consumption is expressed as:

$$\begin{aligned}\mathbf{E}(\mathcal{Y},\mathcal{Z},\mathbb{P},\mathbf{P_{active}},\mathbf{P_{sleep}}) &= \sum_{m\in\mathbb{B}_\mathbb{A}} (P^m_{active}\cdot z_m) \\ &+ \sum_{f\in\mathbb{B}_\mathbb{O}} [P^f_{sleep}\cdot(1-z_f)] + \sum_{s\in\mathbb{B}_\mathbb{S},l\in\mathbb{U}_\mathbb{L},k\in\mathbb{Q}_\mathbb{K}} (p^s_{lk}\cdot y^s_{lk}) \\ &+ \sum_{f\in\mathbb{B}_\mathbb{O}} (P^f_{active}\cdot z_f) + \sum_{m\in\mathbb{B}_\mathbb{A}} [P^m_{sleep}\cdot(1-z_m)].\end{aligned}$$

To fulfill the requirement, BSs chooses MCS, delivering data at a necessary level of transmit power over a sufficient number of resource blocks. The model is designed for finding the optimal operation policy for BSs by deciding y^s_{lk} and z_s. When y^s_{lk} in the solution is equal to 1, user l is associated with BS s through modulation technique k, which corresponds to r_{lk} resource blocks and p^s_{lk} to satisfy the requirement i_l.

The formulation for the BS operation and user association assignment problem can be expressed as follows:

$$\begin{aligned}\max \quad & \frac{\mathbf{S}(\mathcal{Y},\mathbb{I},\mathbb{D})}{\mathbf{E}(\mathcal{Y},\mathcal{Z},\mathbb{P},\mathbf{P_{active}},\mathbf{P_{sleep}})} \\ \text{s.t.} \quad & \text{Constraints}\,(1)-(9), \\ & y^s_{lk}, z_s \in \{0,1\}, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K}.\end{aligned} \tag{10}$$

The problem (10) has the objective function to maximize the spectrum-energy efficiency. Note that due to the energy required for sleep mode, the energy consumption function $\mathbf{E}(\cdot)$ has a lower bound even though no associations is formed by BSs. The case, therefore, will not arise that total energy consumption is very small (e.g., goes to zero). However, the problem involves a fractional objective function which we reformulate into a MILP problem to obtain the optimal solution.

4.1 Problem Transformation

In (10), although the constraints are linear, the objective function is a ratio of two linear terms, hence making the model nonlinear. To eliminate the nonlinearity, the objective function must be transformed to be pure linear; also when there exists a solution to the transformed model, a solution can also be found to the original model. The objective function can be transformed to a linear function as follows. Since in the original problem, the denominator is always positive over the feasible sets of $\mathcal{Y}$ and $\mathcal{Z}$, variables μ^s_{lk}, ν_s and τ are introduced, which hold $\mu^s_{lk} = \tau \cdot y^s_{lk}$ and $\nu_s = \tau \cdot z_s$, and $\tau = \frac{1}{\mathbf{E}(\cdot)}$. The transformed problem is presented below:

$$
\begin{aligned}
\max \quad & \sum_{s\in\mathbb{B}_\mathbb{S}, l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}} \left(\frac{i_l}{r_{lk}} \cdot \mu^s_{lk}\right) \\
\text{s.t.} \quad & \sum_{m\in\mathbb{B}_\mathbb{A}} [(P^m_{active} - P^m_{sleep}) \cdot \nu_m] + \tau \cdot M \cdot P^m_{sleep} \\
& + \sum_{f\in\mathbb{B}_\mathbb{O}} [(P^f_{active} - P^f_{sleep}) \cdot \nu_f] + \tau \cdot F \cdot P^f_{sleep} \\
& + \sum_{s\in\mathbb{B}_\mathbb{S}, l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}} (p^s_{lk} \cdot \mu^s_{lk}) = 1, \\
& \sum_{s\in\mathbb{B}_\mathbb{S}, k\in\mathbb{Q}_\mathbb{K}} \mu^s_{lk} \le \tau, \forall l \in \mathbb{U}_\mathbb{L} \\
& \sum_{s\in\mathbb{B}_\mathbb{S}, k\in\mathbb{Q}_\mathbb{K}} \mu^s_{ek} = \tau, \forall e \in \mathbb{U}_\mathbb{W} \\
& \sum_{k\in\mathbb{Q}_\mathbb{K}} \mu^s_{lk} \le \tau \cdot c^s_l, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, \\
& \sum_{l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}} (p^m_{lk} \cdot \mu^m_{lk}) \le \tau \cdot P^m_{max}, \forall m \in \mathbb{B}_\mathbb{A} \\
& \sum_{l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}} (p^f_{lk} \cdot \mu^f_{lk}) \le \tau \cdot P^f_{max}, \forall f \in \mathbb{B}_\mathbb{O} \\
& \sum_{\substack{h^m_g\in\mathbf{H_m},\\ l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}}} [r_{lk} \cdot (\mu^m_{lk} + \mu^{h^m_g}_{lk})] \le \tau \cdot R^m, \forall m \\
& \sum_{l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}} \mu^s_{lk} \ge \nu_s, \forall s \in \mathbb{B}_\mathbb{S} \\
& \mu^s_{lk} \le \nu_s, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K} \\
& \sum_{s\in\mathbb{B}_\mathbb{S}, l\in\mathbb{U}_\mathbb{L}, k\in\mathbb{Q}_\mathbb{K}} \mu^s_{lk} \ge \tau \cdot \delta, \\
& \mu^s_{lk}, \nu_s, \tau \ge 0, \forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K}
\end{aligned}
\tag{11}
$$

Provided $\tau > 0$ at the optimal solution, this LP problem is equivalent to the fractional objective problem except the binary condition of decision variables. The values of the variables y^s_{lk} and z_s in the optimal solution to the fractional objective problem are obtained from dividing optimal μ^s_{lk} and ν_s by τ. As decision variables in the fractional objective model are binary, it is necessary to impose on the transformed model the constraints which reflect the binary property. We know if y^s_{lk} is zero, then μ^s_{lk} must be zero; otherwise, $\mu^s_{lk} = \tau$ if $y^s_{lk} = 1$. In the following constraints, binary variables are introduced, α^s_{lk} and β_s, as well as a constant value Q holding a value greater than μ^s_{lk}, ν_s and τ.

$$
\begin{aligned}
\mu^s_{lk} - \tau - Q \cdot \alpha^s_{lk} &\ge -Q, \forall s, l, k \\
\nu_s - \tau - Q \cdot \beta_s &\ge -Q, \forall s, l, k \\
\mu^s_{lk} - \tau + Q \cdot \alpha^s_{lk} &\le Q, \forall s, l, k \\
\nu_s - \tau + Q \cdot \beta_s &\le Q, \forall s, l, k \\
\mu^s_{lk} &\le Q \cdot \alpha^s_{lk}, \forall s, l, k \\
\nu_s &\le Q \cdot \beta_s, \forall s, l, k \\
\alpha^s_{lk}, \beta_s &\in \{0, 1\}, \forall s, l, k
\end{aligned}
$$

Problem (10) is equivalent to (11) given the fact that $y^s_{lk} = \frac{\mu^s_{lk}}{\tau}, z_s = \frac{\nu_s}{\tau}$, $\forall s \in \mathbb{B}_\mathbb{S}, l \in \mathbb{U}_\mathbb{L}, k \in \mathbb{Q}_\mathbb{K}$ and $\tau = \frac{1}{\mathbf{E}(\cdot)}$. The MILP problem considers transmission power, as $\mathbf{E}(\cdot)$ is one of the transformed constraints. Since the the value of τ has the impact on both μ^s_{lk} and ν_s, not only r_{lk} but also p^s_{lk} influences the objective value.

5 Approaches to Solve the Problem

Software packages, CPLEX for instance, are able to solve MILP problems; however, a major disadvantage of MILP is its computational complexity. Because MILP is NP-hard in general, computational requirements can grow significantly as the number of binary variables increases. Motivated to generate less computationally intensive methods, we develop two techniques solving the formulated problem by finding suboptimal solutions, which are discussed in the following.

Remark 1. *The formulated problem is NP-hard by reduction from the set-packing problem, which is a well-known NP-complete problem.*

Proof. In the set packing optimization problem, we need to find a set packing that uses the most sets for maximizing the total value. Such problem can be defined as follows. Given a universal set V and a set F which consists of some subsets of V, a packing is a subset $G \subset F$ of sets such that all sets in G are pairwise disjoint.

To obtain the reduction from the set-packing problem, we consider a simplified version of the formulated problem where a HetNet consists of a single macro, a single small cell, and N UEs uniformly distributed within the coverage area. We first set the overall energy consumption as a constant so that the objective function is linear. Suppose there is only one modulation scheme, and resource blocks and transmit power are always sufficient. For both BSs, the operation mode is active, and all users are associated either of them. For UEs,

Algorithm 1: $\max_x f(x)$.

Given $l \leq f^* \leq h$ and the tolerance $\varepsilon > 0$;
while $(h - l) > \varepsilon$ **do**
  $\alpha = (h + l)/2$;
  Solve the feasibility problem (12);
  if *feasible* **then** $l = \alpha$;
  else $h = \alpha$;
end

the transmit power is identical from BSs. In this case, the overall energy consumption is constant, and we maximize spectrum efficiency as optimizing spectrum-energy efficiency.

Let $U_L = V$, and F can be constructed based on the topological relation constraint. For $F_s \subset F$, F_s is a set that represents the possible sets containing $v \in V$ as long as UE v is in the coverage of BS s. The problem now becomes to determining UEs to cover by choosing the sets from F such that spectrum efficiency is maximized. Thus, the set packing problem is a special case of our problem. It is straightforward that the reduction is in polynomial time, and the set-packing problem has a solution if and only if the constructed problem has a solution. Since the set-packing problem is a well-known NP-complete problem, it follows that our problem is NP-hard. □

5.1 Quasiconvex Optimization Problem

A quasiconvex problem takes the form:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & f_k(x) \leq 0, k = 1, \ldots, m, \\ & \mathbf{Ax} = \mathbf{b}. \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is quasiconvex, $f_1, \ldots, f_m$ are convex, and $\mathbf{a}_k \in \mathbb{R}^n$ and $b_k \in \mathbb{R}$ for $k = 1, \ldots, p$, are affine. The objective function in (10) is quasiconvex according to Definition 1:

Definition 1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is ***quasiconvex*** if its domain $dom(f)$ and all its sublevel sets S_α are convex,

$$S_\alpha = \{x \in dom(f) | f(x) \leq \alpha\}, \forall \alpha \in \mathbb{R}.$$

Proof. Let the function

$$\frac{\mathbf{S}(\mathcal{Y}, \mathbb{I}, \mathbb{D})}{\mathbf{E}(\mathcal{Y}, \mathcal{Z}, \mathbb{P}, P^m_{active}, P^f_{sleep})} = f(x) = \frac{a^T x + b}{c^T x + d}.$$

As $dom(f) = \{x | c^T x + d > 0\}$ and its α-sublevel set is

$$\begin{aligned} S_\alpha &= \{x | c^T x + d > 0, (a^T x + b)/(c^T x + d) \leq \alpha\} \\ &= \{x | c^T x + d > 0, (a^T x + b) \leq \alpha(c^T x + d)\}, \end{aligned}$$

which is convex. Therefore, $f(x)$ is quasiconvex. □

The inequality constraints are convex because they are linear and the equality constraint is affine since the solution set of a linear equation is an affine set.

Algorithm 2: Spectrum-Energy Efficiency Maximization

Input: $S, L, K, M, F, N, E, R_b, \mathbb{B}_\mathbb{S}, \mathbb{U}_\mathbb{L}, \mathbb{Q}_\mathbb{K}, \mathbb{B}_\mathbb{A}, \mathbb{B}_\mathbb{O}, \mathbb{U}_\mathbb{X}, \mathbb{U}_\mathbb{W}, \mathbb{C}, \mathbb{I}, \mathbb{D}, \mathbb{P}, \mathbf{H_m}, \mathbf{R}, \mathbf{P_{active}}, \mathbf{P_{sleep}}$ and $\mathbf{P_{max}}$.
Output: $\mathcal{Y}$ and $\mathcal{Z}$.
Compute $g^s_{lk} \in \Omega, \forall s, l, k$;
for $l = N + 1$ to L **do**
  Find s and k combination that gives best g^s_{lk};
  $y^s_{lk} \leftarrow 1$ and $z_s \leftarrow 1$;
for $s = 1$ to M **do**
  Macro-Femto();
Sleeping();
while *there exists an* s *that can cover more users* **do**
  Zooming();
for l *having no service* **and** s *active* **do**
  if *exists* s, l, k *combination to improve* $\mathcal{C}$ **then**
    $y^s_{lk} \leftarrow 1$ and $z_s \leftarrow 1$;
while *constraint* (9) *is not satisfied* **do**
  Find $s, \forall z_s = 0$ covering the most non-served users;
  for $c^s_l = 1$ **do**
    if $y^s_{lk} \leftarrow 1$ *satisfies constraints* (4)-(6) **then**
      $y^s_{lk} \leftarrow 1$ and $z_s \leftarrow 1$;
for l *having association* **do**
  if there exists a k that can improve $\mathcal{C}$ **then**
    $y^s_{lk} \leftarrow 1$ and $z_s \leftarrow 1$;
return $\mathcal{Y}$ and $\mathcal{Z}$.

Hence (ObjectiveFunction) is a quasiconvex optimization problem, whose global optimum can be computed via a sequence of convex feasibility problems. Let f^* denote the optimal value of the quasiconvex object function. Given $\gamma \in \mathbb{R}$, if the convex feasibility problem of (12) is feasible, then we have $f^* \geq \gamma$.

$$\begin{aligned} \text{find} \quad & x \\ \text{s.t.} \quad & f(x) \geq \gamma, \\ & f_k(x) \leq 0, k = 1, \ldots, m, \\ & \mathbf{Ax} = \mathbf{b}. \end{aligned} \tag{12}$$

Conversely, if the above problem is infeasible, then we can conclude $f^* < \gamma$. Thus we can check whether the optimal value f^* is less or more than a given value γ. Algorithm 1 shows the procedure of the bisection method for solving the problem. It starts with a range $[l, h]$ that is known to contain f^*. Then we solve the feasibility problem at its mid-point $\alpha = (l + h)/2$. Depending on whether it is feasible, the algorithm continues on the identified half of the interval. The algorithm is guaranteed to converge in $\lceil \log_2((h - l)/\varepsilon) \rceil$ iterations. The bisection algorithm is less expensive in terms of computation complexity.

5.2 Heuristic Algorithm

In the wireless communication environment the number of users changes and the requirements fluctuate

frequently; hence, heuristic algorithms are favored to solve the proposed optimization model in real time. The approach is shown in Algorithm 2, which accepts same input. While the determination of operation and association depends on information that changes in each frame, thereby tackling the maximization problem within the per-frame optimized decision, the algorithm can perform maximization at coarse time scales if per-frame efficiency maximization is infeasible due to practical constraints. Since the ongoing connections are guaranteed to be maintained, the situation is prevented where BSs are switched on and off between solutions. BSs report dynamic parameters to facilitate the solution: (i) $\mathbb{U}_{\mathbb{X}}$, $\mathbb{U}_{\mathbb{W}}$, $\mathbb{I}$ ($\mathbb{D}$, $\mathbb{P}$ are then estimated with each MCS through the screening procedure), and (ii) $\mathbb{C}$ through the ranging process. Other input of Algorithm 2 is static and determined with the network deployment by operators. Although capable of reducing computation time, when the heuristic approach is applied to an extremely large-size network, the solution might not be able to obtained in the time of a frame. In this regard, tuning the algorithm by further reducing its complexity can be a fix.

A preliminary solution is initialized by, for each existing user, selecting the cell and modulation technique which in combination allow the best single link spectrum-energy efficiency. The solution is firstly updated in the manner of migrating active users from macro cells to femto ones, looking for the opportunity of sleeping macro BSs. Then, the heuristics seeks for favorable possibilities under which users can be reassociated with the active cells other than current ones. If all users are swept into its neighboring BSs, the cell can be switched to sleep mode. Subsequently, there may exist users who are not served. Without activating BSs in a sleep state, assigning these users to awake BSs could benefit improving system spectrum-energy efficiency. On the other hand, the constraint on the served user number might not be presently fulfilled, which is responded to as inactive BSs encompassing the most non-served users will be turned on, accommodating the users with no association. Lastly, a better solution can be achieved through changing modulation technique to live links for improving system spectrum-energy efficiency. Overall, the computation complexity of the algorithm is $\mathcal{O}(SL^3K)$.

5.3 Algorithm Description

The spectrum-energy efficiency of a single link ($g^s_{lk} = \frac{i_L}{\frac{r_{lk}}{p^s_{lk}}}$) is computed for each combination of BS, user and modulation technique. An initial solution is suggested with existing users to choose best $g^s_{lk}, \forall s \in \mathbb{B}_{\mathbb{S}}, k \in \mathbb{Q}_{\mathbb{K}}$. **Macro-Femto**() looks into active macro BSs in order for examining the possibility of user migration to femto cells facilitating sleeping macro cells. During **Macro-Femto**(), associated users with the macro cell are considered for being migrated to femto cells. Two conditions are necessary for getting into a sleep state: all the users are able to be migrated and energy consumption can be reduced. Running many femto cells may consume energy more than that what can be reduced from sleep macro BSs, which is not beneficial to spectrum-energy efficiency improvement. Therefore, starting from the macro cell that could have the highest migrated user number, **Sleeping**() chooses femto cells that require less amount of energy to sleep. During the procedure, if the total transmist power of a femto BS is higher than $P^f_{max}, \forall f \in \mathbb{B}_{\mathbb{O}}$, modulation technique will be adjusted to reduce the imposed energy, from users who need the highest transmitting power.

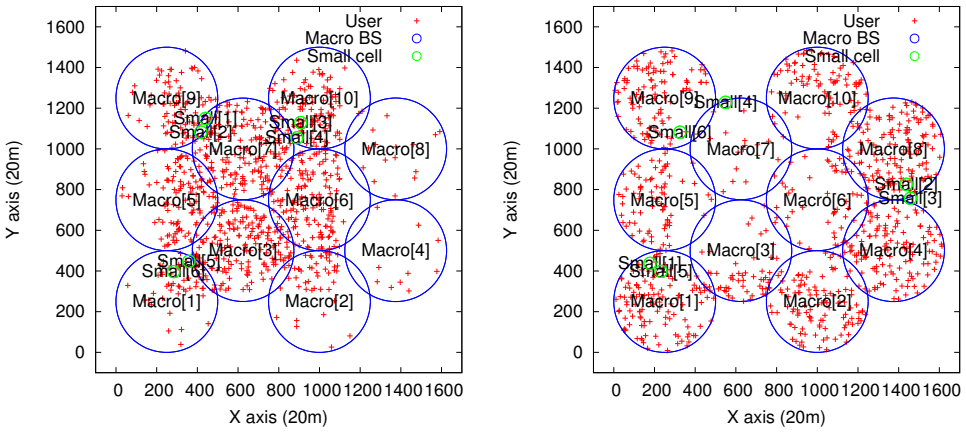
Moreover, it is with the intention of economizing on energy consumption by engaging fewer BSs in active mode. Starting from the one showing most active users in coverage, active BSs connect users who were served by others as many as they can, providing constraints (4)-(6) are not violated, which essentially is what **Zooming**() for. In the case of the insufficient resource block number or total power to deploy, appropriate modulation adjustment is considered. Specifically, depending on which resource is not sufficient, for users taking high transmit power or large resource blocks, modulation is tuned till no more adjustments can or is needed to be made. For no-service users, two active BSs are chosen to bring two highest g^s_{lk}. From the user giving best g^s_{lk}, if efficiency can be improved, users pick first choices in the first iteration; if the total served user number cannot be accepted, second choices come to play for the rest with no association. Until the blocking ratio constraint is met, BSs are turned on and provide the users services via best modulation method. Then the solution is updated if spectrum-energy efficiency can be enhanced by modulation adjustment.

TABLE 3
System Parameters

Parameter	Value
Channel bandwidth	10 MHz, frequency reuse 3
Sleep mode power	$P^m_{sleep} = 8, P^f_{sleep} = 0.028$ (W)
Active mode power	$P^m_{active} = 500, P^f_{active} = 5$ (W)
Maximum transmission power	$P^m_{max} = 40, P^f_{max} = 0.14$ (W)
Required data rate	Random (0-9.9 Mbps per user)
User arrival rate	λ =30 per macrocell
Resource block number	600 per DL subframe

6 Numerical Results

This section presents case studies where the wireless wide area network (WWAN) is assumed such as LTE Advanced using OFDMA interface for the downlink. The case studies are conducted to evaluate the effectiveness of the formulated mathematical model and proposed heuristic algorithm in terms of produced solutions, and the computational efficiencies in multi-cell-multi-users scenarios. We firstly evaluate performance improvements of the proposed solution, and then compare our scheme with others to show the effectiveness. The computation cost and solution obtained by the



(a) Type 1 distribution

(b) Type 2 distribution

Fig. 2. Illustration of node distribution setup.

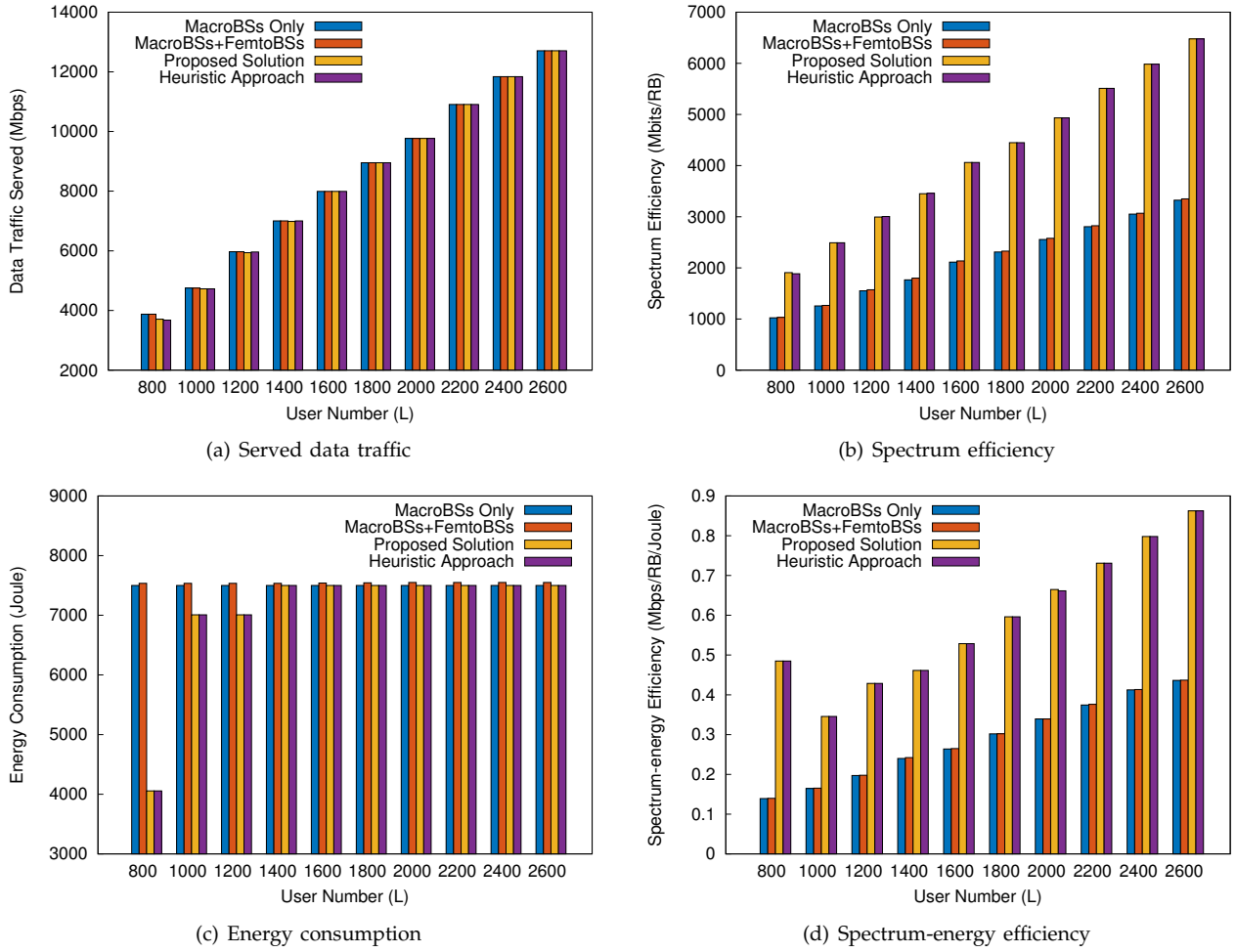
heuristic algorithm are studied as well. The main system parameters taken into account in the simulations are tabulated in Table 3 unless stated otherwise.

As WWAN spans a relatively large area, we separate areas of coverage into three geographic areas, urban (central), rural (peripheral) and suburban (in-between) regions. The building penetration loss correction values vary in these areas, which are 12, 18 and 15 dB, respectively [22]; the area percentages are 65%, 15% and 20%, respectively. Three types of user distribution are considered across the entire area. Fig. 2(a) shows the illustrative Type 1 network layout, in which 80% of the users are in the urban area, 12.5% suburban and 7.5% rural, respectively; Type 2 distribution is plotted in Fig. 2(b), where the peripheral region has a larger population. The other type of user distribution is uniform, for which user numbers are basically the same in the three areas. In our study, the frequency reuse 3 scheme is adopted on macro BSs and the resource block constraint is expressed as Eq. (4); for each macrocellular BS, it shares the reused radio resource with its small cells. The constraint can be modified in order to work with any given spectrum allocation policy where frequency bands are allocated individually. For instance, in split spectrum macrocell-femtocell, certain portion of the available spectrum resources are dedicated to each tier to avoid interference problems among the tiers.

6.1 Solving the Model

Fig. 3 shows the results obtained from four schemes for cases with user numbers from 800 to 2600 and Type 1 distribution. There are 15 macro and 10 femto BSs, and the blocking ratio is 6%. The four schemes includes homogeneous macrocellular network, heterogeneous network consisting of both macrocell and femto cell sites, proposed framework that optimizes spectrum-energy efficiency of HetNets, and heuristic approach that solves the formulated problem (Algorithm 2), respectively. In Fig. 3(a) first two schemes serve all users, and therefore fulfils more data requirements, whereas the proposed solution is allowed to block users; hence in the cases with numbers from 800 to 1400, it serves less. For other cases, since serving more users improves spectrum-energy efficiency, all users are served. In Fig. 3(b), with optimal modulation selection, the proposed scheme obtains the highest spectrum efficiency for all cases, which is on average improved by 92% from HetNet scenarios, while HetNets outperform homogeneous macrocellular networks slightly. Note that the improvement in spectrum efficiency correlates with small cell user number. We run additional simulations and the results show as more small cells are introduced, the improvement in spectrum efficiency increases. Since the possibility of being associated with small cells increases, we can expect the increase in the percentage of users associated through better MCS.

Fig. 3(c) shows the total energy consumed by different schemes in different cases. For the homogeneous network scheme, without turning off BSs, all BSs stay active even in unsaturated networks (i.e. $L < 1200$), resulting in heavy overall energy consumption, whereas the HetNet scheme consumes more energy than because femtos account for the additional energy use for transmission and being active. For our scheme, when 800 users are in the network, almost half of the BSs are switch off, saving more than 45% energy in comparison with that of the HetNet scheme. But the difference becomes to around 7% when the user number grows to 1000 and 1200; all macro cells stay active to serve more than 1400 users.As the number of users increases, the energy consumption difference decreases between the HetNet scheme and the proposed scheme due to the decreased number of BSs that can operate in sleep mode. For example, while 1200 users are considered, 6 more macro BSs remain active compared with the case of 800 users. In the solution, the BSs differ in service range based on the distribution of users and their traffic service requirements through adjusting the modulation scheme and transmit power,



(a) Served data traffic

(b) Spectrum efficiency

(c) Energy consumption

(d) Spectrum-energy efficiency

Fig. 3. Performance comparisons among four schemes.

carrying out cell zooming in/out. The BSs with no users to serve can be turned off, applying BS sleep mode; the users may connect with another BSs, realizing migration.

Fig. 3(d) depicts the performance improvement on spectrum-energy efficiency. It is shown that the spectrum-energy efficiency is increased slightly in the HetNet scheme from the homogeneous scheme, whereas the efficiency capacity improves substantially in the proposed scheme with an average improvement of 114%. When the user number is 800, because a significant number of BSs are in sleep mode, the efficiency is enhanced by almost 250%. Spectrum-energy efficiency takes into account energy efficiency and spectrum efficiency together. By maximizing this efficiency capacity, both metrics can be improved. Note that optimizing either metric does not necessarily achieve optimal spectrum-energy efficiency due to the trade-off.

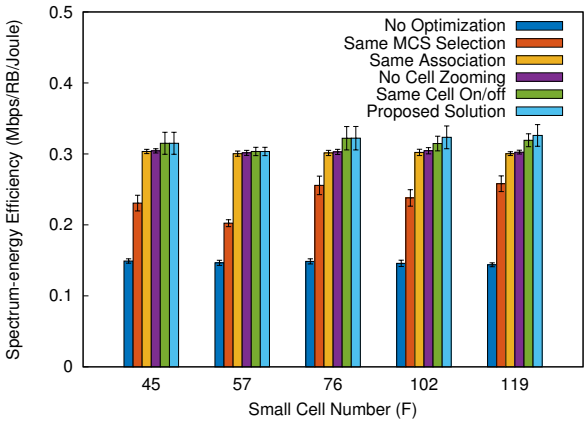


Fig. 4. The Proposed Scheme versus Baseline Schemes.

6.2 Comparative Schemes

We next compare the suggested scheme against six different configurations. They are (i) no optimization is realized upon the HetNet, (ii) the change of MCS selection is not considered ($\mu^{s}_{lk'} = 0, \forall s \in \mathbb{B}_{\mathbb{S}}, l \in \mathbb{U}_{\mathbb{L}}, k' \in \mathbb{Q}_{\mathbb{K}} \setminus \{k\}$, where k is the original MCS selection for users), (iii) optimization does not migrate users ($\mu^{s'}_{lk} = 0, \forall s' \in \mathbb{B}_{\mathbb{S}} \setminus \{s\}, l \in \mathbb{U}_{\mathbb{L}}, k \in \mathbb{Q}_{\mathbb{K}}$, where s is the original BS user l is associated with), (iv) BSs do not zoom out to cover neighboring users; hence, coverage will not be extended (effecting Eq. (3)), (v) the BSs with no users

TABLE 4
Spectrum-energy Efficiency Improvement in Scenarios

Number of Nodes			User Distribution	Spectrum-Energy Efficiency Improved
MacroBS	FemtoBS	User		
15	25	4000	Type 1	95.39%
			Type 2	99.43%
			Uniform	95.48%
18	30	5000	Type 1	93.30%
			Type 2	96.54%
			Uniform	94.03%
18	35	6000	Type 1	94.07%
			Type 2	96.81%
			Uniform	94.99%
20	40	7000	Type 1	91.51%
			Type 2	97.54%
			Uniform	95.06%
20	45	8000	Type 1	92.84%
			Type 2	97.62%
			Uniform	95.71%

to serve originally are not considered to associate with any in optimal solutions ($\nu'_s = 0, \forall s' \in \mathbb{B}_{\mathbb{S}}$, where s' is BS has no users within its original coverage), and (vi) the proposed maximization scheme is applied. Except the first configuration, the spectrum-energy capacity is optimized for all others.

Fig. 4 shows our results. Without optimizing MCS, the spectrum-energy capacity drops by on average 33%, which means the most sensitive dependence is on MCS selection, compared with the other three key configurations. The reason is optimizing the MCS selection picks better MCS values than the default MCS in the sense of balancing overall energy consumption and resources (e.g. higher power but less RBs). We did the experiment to verify that the capacity difference becomes greater with more UEs, due to more MCS selection improvement opportunities. Furthermore, in Fig. 4, since migration operation is excluded, the existing mobile users cannot be re-associated with other BSs, resulting in more BSs that are needed to be activated; therefore 5.5% decrease from the proposed configuration. Similarly, the differential between default cell zooming and the proposed solution averages out 5%, for which the main reason is without the ability to expand coverage areas, the chances are reduced that the serving users are taken over by other BSs and the BS is allowed to be in sleep mode. Lastly, the performance of the default cell on/off configuration depends on the distribution of UEs. When no UEs are in the coverage of femtos, they are switched off; when more UEs appear in the area of femtos, they cannot sleep.

Table 4 presents improvements in spectrum-energy efficiency for various numbers of nodes with different user distribution considerations. It is manifested that the optimization model is capable of being scaled, as the spectrum-energy efficiency can be improved by more than 91% for a variety of network settings regardless of the number of nodes and user distribution. For operators, it is economically essential to be concerned about how efficiently resource blocks are utilized and energy is consumed. The formulated model looks for assignments (BSs, users and modulation/coding schemes) which together contribute the best spectrum-energy efficiency.

6.3 Solution with Other Spectrum Partitioning Model

To show that our formulation can adopt other spectrum sharing model between macro and femto cells, we also evaluate our algorithms considering small cells resources are allocated using frequency reuse of 1. In this subsection we incorporate our model with Almost Blank Sub-frames (ABS periods), which are a recently proposed and discussed technique by the 3GPP project [24]. ABS periods are certain periods designated for each macro to remain silent, over which the femto can use for downlink transmissions, improving performance to users close to the femto. We modify the proposed formulation to jointly optimize ABS parameters for each cell. The goal of our evaluation is to compare our proposed solution with other alternative schemes.

We follow the operational LTE network setup in [25] to evaluate our algorithms in the context where ABS is used. An area of around 8.9 km^2 is selected which has a 28 macro BSs and 10 femto BSs. The user locations are uniformly chosen for our evaluation under the following rules. In the area under consideration, we chose a user density of around 190 users/sq-km (urban). In addition, we create user hotspots around 4 femtos which have double the user density. One other femto has 50% more traffic. We also perform evaluation by varying the user density around the macro cells to 95 users/sq-km (suburban) and 50 users/sq-km (rural) without altering the hot-spot user densities around the selected femtos.

6.3.1 *Optimal ABS parameters*

For fair comparison, along with ABS patterns, we consider as well the concept of cell selection bias (CSB), allowing users to bias its association towards the small cell by a margin; therefore extending cell range [26]. We compare our algorithm to the schemes with the following two (ABS, CSB) combinations: (1/8, 0 dBm) and (2/8, 5 dBm). We also compare ours with the scheme where based on the user association from fixed ABS and CSB pattern, the modulation scheme selection is optimized to improve spectrum-energy efficiency. Furthermore, a 2-step local optimal based scheme is implemented. This scheme works as follows. First, each femto maximizes the total improvement of spectrum-energy efficiency with serving all users within the coverage range of the femto through best modulation schemes from users. This step readily provides the solution on users that associate with femtos. In the next step, each macro m obtains the fraction (for example, x_m) of RBs occupied by femtos within its coverage range and then macro offers $\lceil R^m x_m \rceil$ as ABS-sub-frames. Each femto can only use minimum number of ABS-sub-frames offered by its macros. Then,

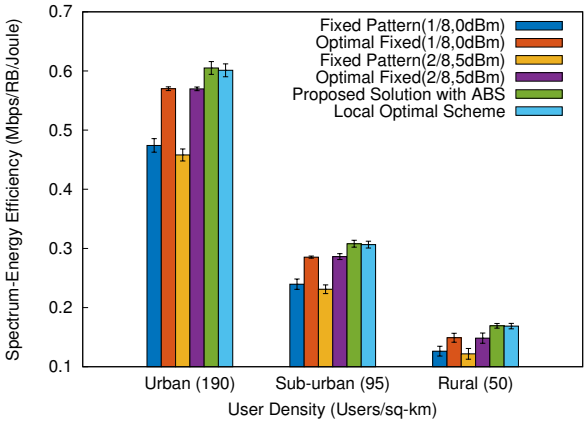


Fig. 5. Comparison of ABS pattern.

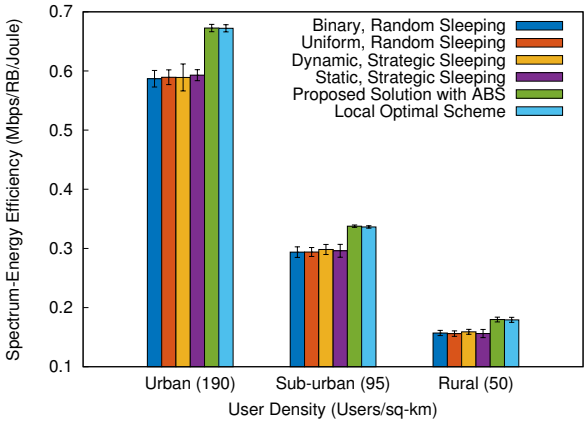


Fig. 6. Comparison of sleeping strategy.

with the decision on femtos, the decision is made by optimization on the remaining users that associate with macros. For our results, we consider serving all UEs with a blocking ratio of 0.

In Fig. 5, we compare our algorithm to different network wide fixed ABS settings. The fixed (ABS, CSB) setting of (2/8, 5 dBm) performs the worst due to more activated stations. With CSB, the problem of under-utilization of femto cells may be eased, since more users may be offloaded to femtos. However, users within the coverage area of femtos could end up being forced to associate with femtos which can potentially be turned off, resulting in more stations that remain in operation. This shows finding a good but fixed cell range setting could be challenging. On the other hand, with optimal modulation scheme selection, efficiency can be improved from fixed configuration as shown from the optimal fixed schemes in the figure.

However, the optimal fixed schemes fail to account for the overall active BS number as they only optimize modulation selection, whereas the local optimal scheme and the proposed algorithm do not necessarily activate all BSs in the interested area. Our scheme outperforms the local optimal one since it further reduces the number of active femtos in the network; although it is small for the considered network deployment, the margin can increase substantially in the case of dense deployment of small cells, which is a typical development in practice. On the other hand, the local optimal scheme is easy to implement and could be promising with additional minor changes for a lower-complexity solution.

6.3.2 Base station sleep mode strategies

In this subsection, we compare with other schemes that minimize the power consumption at each BS. Similar to [4], [27], we implement polices of random and strategic sleeping to switch off BS. In *random sleeping*, we model the sleeping strategy as a Bernoulli trial such that each station remains in operation with probability q, independently of all the others. Instead of randomly switching BSs off, we can switch off BSs when their activity levels are low. Specifically, we model the *strategic sleeping* as a function $s(\cdot)$. If the activity level of the coverage area associated with the BS has activity level x, then it continues to operate with probability $s(x)$, independently. We consider two models of random sleeping: *binary* where BSs sleep with probability 0.5, and *uniform* where the probability is drawn from a uniform $[0, 1]$ random variable. We also consider two models of activity levels for the strategic sleeping: *dynamic* where if the activity level in the coverage area associated with the BS is a, then the BS stays awake with probability a, and *static* where if the activity level associated with a BS is greater than a predefined threshold ($\frac{1}{4}$ in our study [28]), then the BS stays awake with probability 1 or it stays awake with probability 0.5 otherwise. We define the activity level as, on each BS, the total number of RBs to serve associated UEs divided by the available RBs for data transmission.

BSs are randomly picked to be activated based on the stay-awake probabilities, and the procedure continues till satisfying the blocking ratio constraint. Then, based on the active/sleep mode decision, user association and modulation scheme selection are optimized. Fig. 6 shows the spectrum-energy efficiency for different schemes, including the sleeping strategies, local optimal scheme and proposed algorithm. The blocking ratio is set to 7.5% and the ABS value is fixed at $\frac{3}{8}$. From the figure, it is shown that at least one of the strategic sleeping strategies is better than random sleeping ones. Although the margin is small, different schemes do not necessarily have a same set of active BSs or same number of active BSs.

We also see that both of proposed algorithm and local optimal scheme have bigger margins of improvement over random and strategic sleeping policies. This is because the four sleeping strategies do not taken into account the option of cooperation among BSs to further minimize energy consumption. Therefore, awake BSs may serve the same areas, which are opportunities for further energy savings. Nevertheless, different from our algorithm, since the local optimal scheme solves the

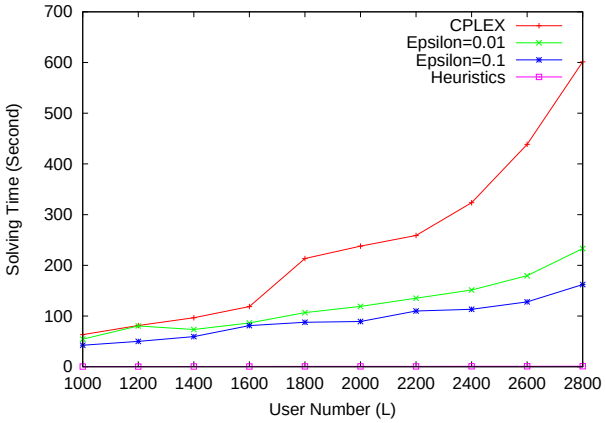


Fig. 7. The solving time comparison.

problem in 2 levels (first femto and then macro), a global optimal solution is difficult to obtain. The figure shows the proposed algorithm performs better for different user densities, suggesting the algorithm is adaptive to fluctuations in traffic activity levels.

6.4 Heuristic Algorithm Results

The formulated model is solvable with CPLEX 12.5, which is a software package dedicated to finding the optimal solution to a MILP problem. The optimal results obtained by CPLEX are taken as the benchmarks to evaluate the proposed heuristic counterparts; the comparisons are measured on a machine equipped with 8 cores. The performance of the devised heuristic approach is shown in Fig. 3. It is indicated that the proposed heuristic algorithm can provide solutions which are very close to the optimum. The proposed algorithm demonstrates computation efficiency as shown in Fig. 7. The solving times of the bisection method are also compared for $\varepsilon = 0.1$ and 0.01, which are the tolerable difference value between the feasible solution with optimality. The computation time of the proposed algorithm increases relatively slightly with the problem size, whereas the computation time shows an exponential growth. In the case of 2800 users, finding the optimal solution spends more than 630 times as the computation time with the heuristics, and bisection methods takes more than 170 times when ε is 10%.

More scenarios are considered to assess the overall effectiveness of the designed algorithm. In Table 5 the problem size grows as numbers of BSs and users increase. The BS number is the summation of macro and femto BS numbers; the computation times of finding the optimal and near-optimal solutions are listed, where solution difference is the difference between the spectrum-energy efficiency values in percentage. The difference made is subtle by the heuristic algorithm. In contrast, the proposed heuristic approach saves 99% of computation times. In addition, we exam the HetNets with a constant number of 18 macro and 30 femto BSs

TABLE 5
Average Solution Difference and Computation Time

Number of Nodes		Solution	Solving Time (Second)	
BS	User	Difference	CPLEX	Heuristics
40	3000	0.0013%	194.98	0.82
40	3500	0.0015%	251.34	1.02
48	4000	0.0015%	437.38	1.26
48	4500	0.0016%	585.58	1.41
53	5000	0.0018%	713.48	1.73
53	5500	0.0021%	841.26	1.92
60	6000	0.0020%	1230.91	2.35
60	6500	0.0020%	1429.50	2.57
65	7000	0.0022%	1679.79	3.03
65	7500	0.0023%	1930.13	3.35

TABLE 6
Gap for Different Topology Types and BS Parameters

		Type 1	Type 2	Uniform
$P^m_{active} = 500$, $P^m_{sleep} = 8$, $P^f_{active} = 5$, $P^f_{sleep} = 0.028$, $P^m_{max} = 40$, $P^f_{max} = 0.14$	$R_b = 0.15$	0.002%	0.304%	0.093%
	$R_b = 0.21$	0.959%	1.681%	0.123%
$P^m_{active} = 130$, $P^m_{sleep} = 75$, $P^f_{active} = 43$, $P^f_{sleep} = 25$, $P^m_{max} = 20$, $P^f_{max} = 6.3$	$R_b = 0.15$	0.006%	0.001%	0.004%
	$R_b = 0.21$	0.007%	0.360%	0.004%

with $\lambda = 36$ and 24. The user number residing in the networks varies from 2600 to 4100 with different user distributions. From Table. 6, the devised heuristics is able to produce spectrum-energy efficiency whose mean difference from optimality is less than 2%. This illustrates that the heuristic algorithm can find effective solutions for the considered network topologies and BS system types. We find that the solutions from our proposed algorithm are fairly close to the optimum for all the considered cases. As the problem size goes up with larger numbers of BSs and users, the computation time rises significantly for obtaining optimal solutions; the heuristic method solves problems more efficiently. The computational efficiency demonstrates scalability of the heuristics in large-scale networks.

7 Conclusion

From mobile service providers' perspectives, saving energy is able to reduce operating expense, and improving spectrum efficiency can increase revenue. To achieve an increase in profitability, it is desired that wireless access networks follow a spectral and energy efficient design. In this paper, we conduct a study on the issue of spectrum-energy efficiency maximization in LTE-A

HetNets, aiming at gaining energy saving and spectrum efficiency improvements. The problem is formulated into the optimization framework that considers the user association, BS operation determination and resources (RBs and energy) allocation. For computational comparison, the bisection approach is adapted to solve the quasi-convex problem, and CPLEX is performed to approach the transformed MILP problem. The heuristic algorithm is developed to suggest a computationally tractable solution to the optimization problem. It is shown in the simulation results that spectrum-energy efficiency is able to be improved significantly from the homogeneous and heterogenous networks in the presented scenarios; also, the optimization problem can be solved efficiently by the proposed algorithm. The established framework lays economic foundation for advanced green wireless networks, providing a guideline for operators in efforts of efficiency and profitability improvements.

REFERENCES

[1] G. P. Fettweis and E. Zimmermann, *ICT energy consumption - trends and challenges*, 2008, no. Wpmc 2008, pp. 2006–2009.

[2] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: a survey," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 59–67, 2008.

[3] H. Claussen, L. T. W. Ho, and L. Samuel, "Financial analysis of a pico-cellular home network deployment," in *Communications, 2007. ICC '07. IEEE International Conference on*, 2007, pp. 5604–5609.

[4] Y. S. Soh, T. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 5, pp. 840–850, May 2013.

[5] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 30 –37, June 2011.

[6] L. Chiaraviglio and P. Torino, "Energy-aware umts access networks," *Scenario*, 2008.

[7] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on*, June 2009, pp. 1 –5.

[8] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 56 –61, June 2011.

[9] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3g cellular networks," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '11. New York, NY, USA: ACM, 2011, pp. 121–132.

[10] J.-M. Kelif, M. Coupechoux, and F. Marache, "Limiting power transmission of green cellular networks: Impact on coverage and capacity," in *Communications (ICC), 2010 IEEE International Conference on*, May 2010, pp. 1 –6.

[11] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *Communications Magazine, IEEE*, vol. 48, no. 11, pp. 74 –79, November 2010.

[12] S. Bhaumik, "Breathe to stay cool : Adjusting cell sizes to reduce energy consumption," *Power*, pp. 41–46, 2010.

[13] B. Badic, T. O'Farrrell, P. Loskot, and J. He, "Energy efficient radio access architectures for green radio: Large versus small cell size deployment," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, September 2009, pp. 1 –5.

[14] F. Richter, A. Fehske, P. Marsch, and G. Fettweis, "Traffic demand and energy efficiency in heterogeneous cellular mobile radio networks," in *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*, May 2010, pp. 1–6.

[15] A. Fehske, F. Richter, and G. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks," in *GLOBECOM Workshops, 2009 IEEE*, Nov 2009, pp. 1–5.

[16] F. Richter, A. Fehske, and G. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *Vehicular Technology Conference Fall, 2009 IEEE 70th*, September 2009, pp. 1 –5.

[17] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "Fluidnet: A flexible cloud-based radio access network for small cells," in *Proceedings of the 19th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '13. New York, NY, USA: ACM, 2013, pp. 99–110.

[18] C.-C. Hsu, J. Chang, Z.-T. Chou, and Z. Abichar, "Optimizing spectrum-energy efficiency in downlink cellular networks," *Mobile Computing, IEEE Transactions on*, vol. 13, no. 9, pp. 2100–2112, Sept 2014.

[19] R. Kwan, C. Leung, and J. Zhang, "Resource allocation in an lte cellular communication system," in *Communications, 2009. ICC '09. IEEE International Conference on*, 2009, pp. 1–5.

[20] H. Klessig, A. Fehske, and G. Fettweis, "Energy efficiency gains in interference-limited heterogeneous cellular mobile radio networks with random micro site deployment," in *Sarnoff Symposium, 2011 34th IEEE*, May 2011, pp. 1 –6.

[21] W. Afric and S. Pilinsky, "Umts lte downlink cell size calculation," in *ELMAR, 2012 Proceedings*, 2012, pp. 105–108.

[22] B. Lannoo, S. Verbrugge, J. Van Ooteghem, B. Quinart, D. Colle, M. Pickavet, P. Demeester, and M. Casteleyn, "Business scenarios for a wimax deployment in belgium," in *Mobile WiMAX Symposium, 2007. IEEE*, March 2007, pp. 132 –137.

[23] J. Penttinen, *The LTE/SAE deployment handbook*. Chichester, U.K: Hoboken, N.J, 2012.

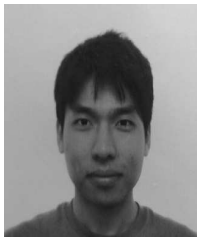
[24] D. Lopez-Perez and H. Claussen, "Duty cycles and load balancing in hetnets with eicic almost blank subframes," in *Personal, Indoor and Mobile Radio Communications (PIMRC Workshops), 2013 IEEE 24th International Symposium on*, Sept 2013, pp. 173–178.

[25] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *IEEE/ACM Trans. Netw.*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[26] A. Tall, Z. Altman, and E. Altman, "Self organizing strategies for enhanced ICIC (eicic)," *CoRR*, vol. abs/1401.2369, 2014.

[27] W. Guo and T. O'Farrell, "Dynamic cell expansion with self-organizing cooperation," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 5, pp. 851–860, May 2013.

[28] H. Chen, Y. Jiang, J. Xu, and H. Hu, "Energy-efficient coordinated scheduling mechanism for cellular communication systems with multiple component carriers," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 5, pp. 959–968, May 2013.



Chan-Ching Hsu received the MS degree in computer engineering from Iowa State University. He is a PhD student in the Department of Electrical and Computer Engineering at Iowa State University. His research interests include multicell processing, green radio, optimization, cognitive and heterogeneous cellular systems, and (beyond) 4G systems.

J. Morris Chang received the PhD degree in computer engineering from North Carolina State University. He is an associate professor at Iowa State University. He received the University Excellence in Teaching Award at Illinois Institute of Technology in 1999. His research interests include wireless networks and computer systems. Currently, he is an editor of the Journal of Microprocessors and Microsystems, and of IEEE IT Professional.