

Energy-Efficient Cell Activation, User Association, and Spectrum Allocation in Heterogeneous Networks

Binnan Zhuang, Dongning Guo, and Michael L. Honig

Abstract—Next generation (5G) cellular networks are expected to be supported by an extensive infrastructure with many-fold increase in the number of cells per unit area compared to today. The total energy consumption of base transceiver stations (BTSs) is an important issue for both economic and environmental reasons. In this paper, an optimization-based framework is proposed for energy-efficient global radio resource management in heterogeneous wireless networks. Specifically, with stochastic arrivals of known rates intended for users, the smallest set of BTSs is activated with jointly optimized user association and spectrum allocation to stabilize the network first and then minimize the delay. The scheme can be carried out periodically on a relatively slow timescale to adapt to aggregate traffic variations and average channel conditions. Numerical results show that the proposed scheme significantly reduces the energy consumption and increases the quality of service compared to existing schemes in the literature.

I. INTRODUCTION

Commercial wireless networks are evolving towards higher frequency reuse by deploying smaller cells to meet increasing demand for mobile data services. In a heterogeneous network (HetNet), macro cells provide for wide area coverage and for serving highly mobile users, whereas dense deployment of femto, pico, and/or micro cells, possibly with distributed antennas, can support much higher data rates per unit area.

The energy consumption due to information and communication technologies worldwide is rising rapidly [1]. As the number of base transceiver stations (BTSs) increases, it is ever more important to manage their energy consumption. An active macro BTS consumes 40 to 80 watts on transmission, whereas the total power consumption is typically hundreds to well over 1,000 watts, which includes the power for signal processing, computation, cooling, and radio frequency power amplification (see, e.g., [2]–[6] and references therein). Hence, savings from transmit power control alone are relatively limited. Much more significant power savings can be accomplished by turning a BTS off or switching it to deep sleep mode.

In this paper, we study how to support given traffic with as few active small cells as possible to conserve energy in a HetNet. Because it may take many seconds to reactivate a BTS in deep sleep [7], the on-off decision should rely on the *aggregate*

traffic and average channels conditions collected over a period, possibly lasting a minute or more. Cell activation/deactivation thus occurs on a much slower timescale than channel-aware scheduling, which typically occurs over time slots of a few milliseconds [8]. The central problem considered here is how to jointly optimize spectrum allocation and user association to minimize the number of active cells. Without loss of generality, we focus on downlink data transmissions.

The problem formulation in this paper builds upon prior work [9]. The following two features distinguish this paper and [9] from most existing work on energy-efficient resource management in the literature, including [10]–[20]. First, we consider stochastic packet arrivals to user traffic queues in lieu of static data rate requirements. Stochastic traffic better models the challenges for small cells as they have much more pronounced traffic variations than macro cells. The proposed slow timescale resource management adapts to the average traffic load, thus avoiding frequent on/off updates due to varying user equipment (UE) data rates. Second, the formulation here facilitates the maximum amount of spectrum agility. Specifically, by considering all possible reuse patterns, arbitrary (possibly nonconsecutive) spectrum can be allocated to a link from any BTS to any UE (see also [21], [22]). This is in contrast to the full-spectrum reuse (i.e., all BTSs use all available spectrum) assumed in [10]–[19].

The optimization problem formulated here is a mixed integer program, where cell activation decisions are expressed as binary variables. The problem is solved numerically using an iterative algorithm based on reweighted ℓ_1 minimization [23]. The method was interpreted as majorization-minimization in [24]. Reference [13] uses the same method to solve the cell activation and user association problem under full-spectrum reuse and static rate requirements.

Numerical results show that the proposed resource management method achieves significant energy savings as well as throughput gains in a typical HetNet. In particular, the performance advantages of spectrum agility is demonstrated by comparing with the method in [13]. As previously noted in [9], spectrum agility is crucial for improving network throughput. Intuitively, improved spectrum allocation allows more BTSs to be turned off. Furthermore, a silent BTS causes no interference to other cells, so that the remaining BTSs may attain higher spectral efficiencies [25]. This may present additional opportunities for turning off BTSs and adding more energy savings.

The proposed method offers a different tradeoff between

B. Zhuang was with the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL, 60208, USA. He is now with Samsung, Inc., San Diego, CA, USA. D. Guo and M. L. Honig are with the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL, 60208, USA.

This work was supported in part by a gift from Futurewei Technologies and by the National Science Foundation under Grant No. CCF-1231828.

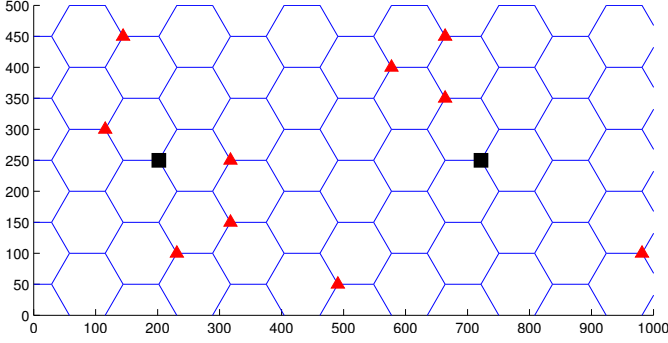


Fig. 1. The topology of a HetNet with $n_1 = 2$ macro BTSs, $n_2 = 10$ randomly deployed pico BTSs and $k = 66$ user groups.

performance and complexity than in [13]. The scheme in [13] is computationally feasible for a HetNet with hundreds of BTSs, whereas the method here is feasible for a cluster of at most 20 to 30 BTSs due to a large number of additional spectrum allocation variables in the optimization problem.

The remainder of this paper is organized as follows. The system model is introduced in Section II. The optimization problems are presented in Section III, and the algorithms are introduced in Section IV. Simulation results are presented in Section V, and conclusions are drawn in Section VI.

II. SYSTEM MODEL

We consider the downlink of a HetNet with n BTSs, including n_1 macro BTSs and $n_2 = n - n_1$ pico BTSs. Denote the set of macro and pico BTSs as $\mathcal{N}_1 = \{1, \dots, n_1\}$ and $\mathcal{N}_2 = \{n_1 + 1, \dots, n\}$, respectively. Let $\mathcal{N} = \{1, \dots, n\}$ denote the set of all the BTSs. All BTSs operate on the same licensed band of W Hz. The frequency resources are assumed to be homogenous on a slow timescale.

The key to total spectrum agility is the notion of *pattern* [9]. A pattern over a set of time-frequency resources is a subset of transmitters sharing the resources. In the downlink, a pattern A is a subset of $\{1, \dots, n\}$, and all BTSs in A have simultaneous access to the frequency band associated with the pattern. Assuming known transmit power spectral densities (PSDs), a particular pattern determines the signal-to-interference-plus-noise ratio (SINR) and hence the spectral efficiencies of all links in the network. The allocation problem can then be formulated as how to divide the resources among all 2^n patterns. In a simple example with only 2 BTSs, there are only patterns: $\{1\}$, $\{2\}$, $\{1, 2\}$, and \emptyset , which denote the patterns used respectively by BTS 1 and BTS 2 exclusively, the pattern used by both BTSs, and the pattern used by none, respectively. For every pattern $A \subset \mathcal{N}$, let y_A be the fraction of total bandwidth allocated to it. Clearly,

$$\sum_{A \subset \mathcal{N}} y_A = 1, \quad (1)$$

and any efficient allocation would set $y_\emptyset = 0$.

At the slow timescale considered it is reasonable to treat users near each other with similar quality of service (QoS) requirements as a *group*. Denote the set of all user groups as $\mathcal{K} = \{1, \dots, k\}$. An example HetNet model with $n = 12$

BTSs and $k = 66$ user groups is shown in Fig. 1, where each group of UEs is assumed to be located at the center of each hexagon (as in [25]). For ease of characterizing the delay as the objective, the aggregate traffic of group j UEs is modeled by Poisson traffic arrivals with arrival rate λ_j . The packet length is exponentially distributed with average length L . It is possible to adopt a different traffic and queueing model (see, e.g., [26]).

We assume each BTS assigns different spectrum resources to different user groups. This can be viewed as statistically multiplexing the packet streams from different user groups. The packets from the same group of UEs are served according to a ‘first in first out’ policy. Hence, each user group effectively has a virtual queue. We also assume multiple BTSs can serve the same group of UEs.¹

We assume BTS i transmits to its UEs with fixed flat transmit PSD, p_i . The spectral efficiency of the link $i \rightarrow j$ under pattern A depends on the receive power and the interference. For concreteness in obtaining numerical results, we use Shannon’s formula to obtain:

$$s_A^{i \rightarrow j} = \frac{W \mathbb{1}(i \in A)}{L} \log_2 \left(1 + \frac{p_i g^{i \rightarrow j}}{\sum_{i' \in A \setminus \{i\}} p_{i'} g^{i' \rightarrow j} + n_j} \right) \quad (2)$$

in packets/second, where $\mathbb{1}(i \in A) = 1$ if $i \in A$ and $\mathbb{1}(i \notin A) = 0$ otherwise, $g^{i \rightarrow j}$ is the power gain of the link $i \rightarrow j$, and n_j is the noise PSD at group j UEs. The link gain $g^{i \rightarrow j}$ includes pathloss and shadowing effects over the slow timescale considered in this paper. Hence $g^{i \rightarrow j}$ is a constant in each decision period independent of the frequency. If small scale fading is included in $g^{i \rightarrow j}$ on the slow timescale, then ergodic spectral efficiency must be used instead of (2), since the decision period spans many coherence time intervals.

All packets intended for group j UEs arrive at an M/M/1 queue. Denote the bandwidth allocated to BTS i to serve group j UEs under pattern A as $x_A^{i \rightarrow j}$. The service rate for this queue contributed by BTS i under pattern A is $s_A^{i \rightarrow j} x_A^{i \rightarrow j}$. This rate is guaranteed regardless of the activities of the other BTSs and queues. The total service rate for queue j is a linear function of the bandwidths:

$$r_j = \sum_{i \in \mathcal{N}} \sum_{A \subset \mathcal{N}} s_A^{i \rightarrow j} x_A^{i \rightarrow j} \quad \text{packets/second.} \quad (3)$$

Hence the average packet sojourn time of the M/M/1 queue for group j UEs takes a simple form [27]:

$$t_j = \frac{1}{(r_j - \lambda_j)^+} \quad \text{seconds,} \quad (4)$$

where $(x)^+ = \max\{0, x\}$, so $t_j = +\infty$ if $r_j \leq \lambda_j$.

¹On a slow timescale this can be realized by letting different BTSs serve different subsets of individual UEs.

III. OPTIMIZATION FRAMEWORK

A. Baseline Formulation

The basic energy-efficient resource allocation problem is formulated as:

$$\underset{\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z}}{\text{minimize}} \quad \sum_{i \in \mathcal{N}_2} c_i z_i \quad (\text{P1a})$$

$$\text{subject to} \quad z_i \in \{0, 1\}, \quad \forall i \in \mathcal{N}_2 \quad (\text{P1b})$$

$$r_j = \sum_{i \in \mathcal{N}} \sum_{A \subset \mathcal{N}} s_A^{i \rightarrow j} x_A^{i \rightarrow j}, \quad \forall j \in \mathcal{K} \quad (\text{P1c})$$

$$\sum_{j \in \mathcal{K}} x_A^{i \rightarrow j} \leq y_A, \quad \forall i \in \mathcal{N}, A \subset \mathcal{N} \quad (\text{P1d})$$

$$\sum_{A \subset \mathcal{N}} \sum_{j \in \mathcal{K}} x_A^{i \rightarrow j} \leq z_i, \quad \forall i \in \mathcal{N}_2 \quad (\text{P1e})$$

$$x_A^{i \rightarrow j} \geq 0, \quad \forall i \in \mathcal{N}, j \in \mathcal{K}, A \subset \mathcal{N} \quad (\text{P1f})$$

$$r_j - \lambda_j \geq \tau_j^{-1}, \quad \forall j \in \mathcal{K} \quad (\text{P1g})$$

$$\sum_{A \subset \mathcal{N}} y_A = 1 \quad (\text{P1h})$$

where:

- The variables \mathbf{r} , \mathbf{x} , \mathbf{y} , and \mathbf{z} are the vector forms of $(r_j)_{j \in \mathcal{K}}$, $(x_A^{i \rightarrow j})_{i \in \mathcal{N}, j \in \mathcal{K}, A \subset \mathcal{N}}$, $(y_A)_{A \subset \mathcal{N}}$, and $(z_i)_{i \in \mathcal{N}_2}$, respectively.
- (P1a) is the total energy cost of the network, where z_i constrained by (P1b) is a binary variable indicating whether pico BTS i is on or off, and c_i is the power cost of pico BTS i if it is active. While the macro BTSs are assumed to be always on to provide basic coverage, it is easy to include their on-off decisions as variables as well.
- (P1c) is the service rate of each user group as given in (3).
- (P1d) guarantees that, for every pattern A , the total bandwidth a BTS allocates to all groups does not exceed the bandwidth assigned to pattern A .
- (P1e) states that a pico BTS i uses no spectrum if it is off and uses at most 1 unit of bandwidth if it is on.
- (P1f) constrains all bandwidths to be nonnegative.
- (P1g) is the QoS constraint for each user group, where τ_j is the maximum delay allowed for group j UEs. The constraint is equivalent to letting t_j in (4) satisfy $0 \leq t_j \leq \tau_j$. In lieu of QoS, quality of experience (QoE) can be considered using the proposed framework as well. To achieve this, each user group can be further divided into smaller groups representing different rate, delay and other QoS preferences. Also, QoE may be modeled a general user-dependent utility function $Q_j(\lambda_j, r_j)$.
- (P1h) constrains the total system bandwidth to be one unit.

The formulation P1 addresses joint user association, spectrum allocation, and BTS activation all under one framework. The objective and all the constraints are linear. Solving P1 minimizes the energy cost by turning off pico BTSs not needed to satisfy the delay requirements.

B. Structure of the Optimal Solution

One concern with P1 is that the spectrum is divided into up to 2^n patterns (segments), which may be impractical for all but

very small networks. Fortunately, we can use Carathéodory's Theorem to show that there exists an optimal allocation that uses no more than k patterns.

Theorem 1: There exists an optimal solution of P1 in which \mathbf{y} is k -sparse, i.e.,

$$\left| \{A \mid y_A > 0, A \subset \mathcal{N}\} \right| \leq k. \quad (5)$$

Proof: We first reformulate P1 by defining a new set of variables $\{v_A^{i \rightarrow j}\}$, which represent the fraction of spectrum under pattern A that BTS i allocates to group j . The constraints (P1c)–(P1f) then become:

$$r_j = \sum_{A \subset \mathcal{N}} \left(\sum_{i \in \mathcal{N}} s_A^{i \rightarrow j} v_A^{i \rightarrow j} \right) y_A, \quad \forall j \in \mathcal{K} \quad (\text{P2a})$$

$$\sum_{j \in \mathcal{K}} v_A^{i \rightarrow j} \leq 1, \quad \forall i \in \mathcal{N}, A \subset \mathcal{N} \quad (\text{P2b})$$

$$\sum_{A \subset \mathcal{N}} \sum_{j \in \mathcal{K}} v_A^{i \rightarrow j} y_A \leq z_i, \quad \forall i \in \mathcal{N}_2 \quad (\text{P2c})$$

$$v_A^{i \rightarrow j} \geq 0, \quad \forall i \in \mathcal{N}, j \in \mathcal{K}, A \subset \mathcal{N}. \quad (\text{P2d})$$

This problem, P2, where $\{v_A^{i \rightarrow j}\}$ replaces $\{x_A^{i \rightarrow j}\}$, is clearly equivalent to P1. Solving P2, the actual spectrum allocations are given by $x_A^{i \rightarrow j} = y_A v_A^{i \rightarrow j}$.

In the remainder of this proof, we show that if a solution $(\mathbf{r}, \mathbf{v}, \mathbf{y}, \mathbf{z})$ to P2 exists, then there exists a k -sparse \mathbf{y}^* and rate tuple \mathbf{r}^* such that $(\mathbf{r}^*, \mathbf{v}, \mathbf{y}^*, \mathbf{z})$ is feasible. This attains the original objective (P1a) because \mathbf{z} remains the same. We shall verify the feasibility constraints pertaining to $(\mathbf{r}^*, \mathbf{y}^*)$, including (P1g), (P1h), (P2a), and (P2c).

Suppose \mathbf{y} has more than k nonzero elements. Let its support be S ($y_A = 0$ if $A \notin S$). Let us define a k -vector \mathbf{t}_A for every $A \in S$ with its elements determined by $t_{j,A} = \sum_{i \in \mathcal{N}} s_A^{i \rightarrow j} v_A^{i \rightarrow j}$. According to (P2a), a convex combination of the vectors $(\mathbf{t}_A)_{A \in S}$ with $(y_A)_{A \in S}$ as coefficients form the optimal rate tuple: $\mathbf{r} = \sum_{A \in S} y_A \mathbf{t}_A$. By Carathéodory's Theorem, \mathbf{r} can be represented as a convex combination of no more than $k + 1$ of those vectors, denoted as $(\mathbf{t}_A)_{A \in S^*}$. Moreover, either \mathbf{r} is on the boundary of the convex hull of $(\mathbf{t}_A)_{A \in S^*}$ or it is an interior point. In either case, there exists \mathbf{r}^* on the boundary that dominates \mathbf{r} in every dimension. Clearly, \mathbf{r}^* is the convex combination of at most k vectors from $(\mathbf{t}_A)_{A \in S^*}$. Therefore, there exists \mathbf{y}^* satisfying (P1h) with support $S^{**} \subset S^* \subset S$, such that $|S^{**}| \leq k$ and

$$\mathbf{r}^* = \sum_{A \in S^{**}} y_A^* \mathbf{t}_A \quad (6)$$

which implies (P2a). Because \mathbf{r}^* dominates \mathbf{r} in every dimension, (P1g) is satisfied. It remains to show that (P2c) holds. For every i with $z_i = 1$, (P2c) is satisfied due to (P1h) and (P2b). For every i with $z_i = 0$, (P2c) requires that $v_A^{i \rightarrow j} = 0$ for every j and every $A \in S$ with $y_A > 0$, which implies that (P2c) remains true for \mathbf{y}^* , because its support is dominated by that of \mathbf{y} . This completes the proof. ■

C. Comparison with Full Spectrum Reuse [13]

The algorithm to be introduced for solving P1 is related to the majorization-minimization approach in [13], which is

based on reweighted ℓ_1 minimization proposed in [23]. To facilitate a fair comparison, we formulate an analogous problem to the one in [13] for cell activation and user association under the current framework. With minor changes, the optimization problem in [13] becomes:

$$\underset{\mathbf{r}, \mathbf{x}, \mathbf{z}}{\text{minimize}} \quad \sum_{i \in \mathcal{N}_2} c_i z_i \quad (\text{P3a})$$

$$\text{subject to } z_i \in \{0, 1\}, \quad \forall i \in \mathcal{N}_2 \quad (\text{P3b})$$

$$r_j = \sum_{i \in \mathcal{N}} s_{\mathcal{N}}^{i \rightarrow j} x^{i \rightarrow j}, \quad \forall j \in \mathcal{K} \quad (\text{P3c})$$

$$\sum_{j \in \mathcal{K}} x^{i \rightarrow j} \leq z_i, \quad \forall i \in \mathcal{N} \quad (\text{P3d})$$

$$x^{i \rightarrow j} \geq 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{K} \quad (\text{P3e})$$

$$r_j - \lambda_j \geq \tau_j^{-1}, \quad \forall j \in \mathcal{K} \quad (\text{P3f})$$

where all variables have the same physical meaning as in P1. The key change from P1 to P3 is that only full spectrum reuse is allowed, where the spectral efficiency of the link $i \rightarrow j$ is $s_{\mathcal{N}}^{i \rightarrow j}$. P3 optimizes user association and the pico BTS on/off selection to minimize the energy cost. The performance of allocations based on P1 and P3 will be compared in Section V using numerical examples.

D. Generalization of the Framework

The objective (P1a) is equal to the weighted ℓ_0 norm of \mathbf{z} :

$$\sum_{i \in \mathcal{N}_2} c_i |z_i|_0 \quad (7)$$

where for any real number z , $|z|_0 = 0$ if $z = 0$ and $|z|_0 = 1$ otherwise. Hence P1 admits an equivalent formulation with the objective function changed to (7) and with (P1b) relaxed to $z_i \in [0, 1]$, $i \in \mathcal{N}_2$.

P1 can also be generalized to utility functions of the form:

$$\sum_{i \in \mathcal{N}_2} c_i |z_i|_0 + f(\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z}) \quad (8)$$

subject to the same constraints, where $z_i \in [0, 1]$ indicates the fraction of the total bandwidth used by BTS i . The objective (8) is highly versatile. For example, letting $f(\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i \in \mathcal{N}} b_i z_i$ accounts for the transmit power consumption being proportional to the bandwidth allocation; letting

$$f(\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z}) = \beta \sum_{j \in \mathcal{K}} \frac{\lambda_j}{\sum_{l \in \mathcal{K}} \lambda_l} \cdot \frac{1}{r_j - \lambda_j}, \quad (9)$$

incorporates the cost of delay in the objective, where β accounts for the tradeoff between energy and delay.

IV. REWEIGHTED ℓ_1 MINIMIZATION

With the weighted ℓ_0 norm (7) as its objective, P1 is a mixed integer program. It is generally difficult to solve due to its combinatorial nature. Such optimization problems frequently appear in sparse signal recovery, portfolio optimization, and statistical estimation. In this paper, we present an algorithm based on a low-complexity method, called *reweighted ℓ_1 minimization* [23].

A. Algorithm Based on Reweighted ℓ_1 Minimization

The basic algorithm for solving P1 consists of iterating between solving a convex optimization problem with weighted ℓ_1 norm relaxation of the ℓ_0 objective (7), and updating the weights. The continuous convex optimization problem based on ℓ_1 relaxation is:

$$\underset{\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z}}{\text{minimize}} \quad \sum_{i \in \mathcal{N}_2} w_i c_i z_i \quad (\text{P4a})$$

$$\text{subject to } z_i \geq 0, \quad i \in \mathcal{N}_2 \quad (\text{P4b})$$

(P1c)–(P1h)

where the objective function becomes a weighted ℓ_1 norm of \mathbf{z} in lieu of the weighted ℓ_0 norm.

Algorithm 1 Reweighted ℓ_1 minimization.

INPUT: $(\lambda_j)_{j \in \mathcal{K}}$, $(s_A^{i \rightarrow j})_{i \in \mathcal{N}, j \in \mathcal{K}, A \subset \mathcal{N}}$, and $(c_i)_{i \in \mathcal{N}_2}$.

OUTPUT: $(\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z})$.

Initialization: $w_i^1 \leftarrow 1$, $\forall i \in \mathcal{N}_2$,

$u^{-1} \leftarrow 0$, $u^0 \leftarrow \sum_{i \in \mathcal{N}_2} c_i$, and $t \leftarrow 1$.

while $t \leq T$ and $|u^{t-1} - u^{t-2}| > \epsilon_1$ **do**

1. Compute $(\mathbf{r}^t, \mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t)$ and the corresponding optimum u^t by solving P4 with $(w_i^t)_{i \in \mathcal{N}_2}$ as the weights.

2. Update the weights by $w_i^{t+1} \leftarrow \frac{1}{z_i^t + \epsilon_2}$, $\forall i \in \mathcal{N}_2$.

3. $t \leftarrow t + 1$.

end while

Algorithm 1 states the iterative procedure. It starts with the weights \mathbf{w} as a vector of all ones. In the t th iteration, the algorithm first computes $(\mathbf{r}^t, \mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t)$ by solving (P4) with \mathbf{w}^t as the weights. In fact, this is a simple linear program. The weights are then updated as $w_i^{t+1} \leftarrow 1/(z_i^t + \epsilon_2)$, where ϵ_2 is some small number. If $\epsilon_2 = 0$, the weight w_i is the inverse of z_i . Hence $w_i z_i$ is a good approximation of $|z_i|_0$. We can also view w_i as a penalty term. If z_i^t is large, then $w_i^{t+1} \approx 0$, whereas if z_i^t is small, then $w_i^{t+1} \gg 1$, so that z_i is more likely to be driven to 0 in the following iteration. The algorithm terminates if either the maximum number of iterations T is reached or convergence (according to some predefined threshold ϵ_1) is achieved.²

An alternative explanation of Algorithm 1 based on the majorization-minimization approach is given in [13]. It makes use of the following property of the ℓ_0 norm [23], [24]:

$$|z|_0 = \lim_{\epsilon \rightarrow 0} g_\epsilon(z) \quad (10)$$

for $z \geq 0$, where

$$g_\epsilon(z) = \frac{\log(1 + z\epsilon^{-1})}{\log(1 + \epsilon^{-1})}. \quad (11)$$

The majorization-minimization method can be regarded as solving a sequence of minimization problems, where in each instance it minimizes a surrogate function that locally majorizes the true objective function (see [24], [28] for more

²Here, convergence refers to the objective not necessarily the variables.

details). For a concave and differentiable function f , a simple and effective majorization-minimization update is:

$$\mathbf{z}^{t+1} = \arg \min_{\mathbf{z} \in \zeta} f(\mathbf{z}^t) + \Delta f(\mathbf{z}^t)^T (\mathbf{z} - \mathbf{z}^t), \quad (12)$$

where ζ is the feasible set of the problem. Since for every $\epsilon > 0$, g_ϵ is concave on $[0, \infty)$, substituting (7) and (10) into (12), we obtain:

$$\mathbf{z}^{t+1} = \arg \min_{\mathbf{z} \in \zeta} \sum_{i \in \mathcal{N}_2} \frac{1}{z_i + \epsilon_2} c_i z_i, \quad (13)$$

where ζ is the feasible set of P4. It is easy to see that (13) is exactly what Algorithm 1 computes in each iteration. Hence Algorithm 1 can be viewed as using the majorization-minimization method to minimize the approximate objective: $f(\mathbf{z}) = \sum_{i \in \mathcal{N}_2} c_i g_{\epsilon_2}(z_i)$. The concavity of $f(\mathbf{z})$ and (13) imply $f(\mathbf{z}^{t+1}) \leq f(\mathbf{z}^t)$, which establishes the convergence of Algorithm 1.

B. A Refined Algorithm

Algorithm 1 mainly deals with the combinatorial nature of P1 introduced by the binary variables \mathbf{z} . However, the number of variables in P4 is $O(kn2^n)$ due to the 2^n frequency patterns. We next reduce the complexity and improve the convergence speed somewhat using the fact that switching each BTS off halves the number of available patterns.

There are two important properties of P4 and Algorithm 1. First, if $z_i = 0$, according to (P1d) and (P1e), then $y_A = 0$ for every A that contains i . Second, if ϵ_2 is close to 0, once Algorithm 1 decides to switch a certain BTS off in some iteration, it is unlikely to be reactivated in later iterations. This is because the weight ϵ_2^{-1} corresponding to $z_i = 0$ is typically much larger than that corresponding to any other nonzero element. Based on these two properties, a refined version of Algorithm 1 is proposed in Algorithm 2, which reduces the dimensions of the feasible set during the iterations.

Algorithm 2 A refined algorithm.

INPUT: $(\lambda_j)_{j \in \mathcal{K}}$, $\left(s_A^{i \rightarrow j}\right)_{i \in \mathcal{N}, j \in \mathcal{K}, A \subset \mathcal{N}}$, and $(c_i)_{i \in \mathcal{N}_2}$.

OUTPUT: $(\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z})$.

Initialization: $w_i^1 \leftarrow 1, \forall i \in \mathcal{N}_2$,

$u^{-1} \leftarrow 0, u^0 \leftarrow \sum_{i \in \mathcal{N}_2} c_i$, and $t \leftarrow 1$.

while $t \leq T$ and $|u^{t-1} - u^{t-2}| > \epsilon_1$ **do**

1. Compute $(\mathbf{r}^t, \mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t)$ and the corresponding optimum u^t by solving P4 with $(w_i^t)_{i \in \mathcal{N}_2}$ as the weights.

2. Update the weights by $w_i^{t+1} \leftarrow \frac{1}{z_i^t + \epsilon_2}$.

if $\exists i \in \mathcal{N}_2$ s.t. $z_i^t = 0$, and $\sum_{i: z_i^t > 0} w_i^{t+1} < \frac{\alpha}{\epsilon_2}$ **then**

$\mathcal{N}_2 \leftarrow \{i \mid z_i^t > 0, i \in \mathcal{N}_2\}$

$\mathcal{N} \leftarrow \mathcal{N}_1 \cup \mathcal{N}_2$

end if

3. $t \leftarrow t + 1$.

end while

The only change from Algorithm 1 is that the BTSs turned off at the end of each iteration are eliminated from future optimizations if a certain condition is met. Namely, if some z_i becomes zero in any iteration, we then compare the sum

of the penalties on all nonzero components, $\sum_{i: z_i > 0} w_i^{t+1}$ against the penalty on any zero component up to a scale factor, namely, α/ϵ_2 . If the total penalty on the nonzero terms is smaller, the pico BTSs with zero z_i will be ignored in future iterations, i.e., future optimizations of P4 will be considered for the reduced pico BTS set $\mathcal{N}_2' = \{i \mid z_i > 0, i \in \mathcal{N}_2\}$. Analogous to Algorithm 1, Algorithm 2 can be regarded as updating a sequence of objective functions f^t in the same form, but with \mathcal{N}_2 updated in each iteration. With slight abuse of notation,³ we have $f^t(\mathbf{z}^t) = f^1(\mathbf{z}^t)$. It is also easy to show that $f^{t+1}(\mathbf{z}^{t+1}) \leq f^{t+1}(\mathbf{z}^t) = f^t(\mathbf{z}^t) \leq f^t(\mathbf{z}^{t-1})$. Hence the values of the approximate objective f^1 after each iteration form a monotonically decreasing sequence, which establishes the convergence of Algorithm 2.

To avoid a premature reduction of the feasible set, the weights on active and inactive BTSs are compared in Algorithm 2. When $\sum_{i: z_i^{t+1} > 0} w_i^{t+1} < \alpha/\epsilon_2$, it is easy to see that an off BTS in the current iteration will not be turned back on in the next iteration according to Algorithm 1. This is because the total reduction of the objective achieved by turning off all currently active BTSs in the next iteration is not enough to compensate for the increased cost due to reactivating a single currently inactive BTS. By setting the scale factor α small enough, the off BTSs that are eliminated from the feasible set are unlikely (albeit still possible) to be turned on again according to Algorithm 1. If l pico BTSs are removed from \mathcal{N}_2 , the number of variables is reduced by a factor of 2^l , which results in a 2^l complexity reduction in future iterations. The performance of Algorithms 1 and 2 will be compared in Section V. Algorithms 1 and 2 can be used to solve P4 if the function f in (P4a) is convex. The only difference then is that the continuous optimization problem in each iteration is a convex optimization problem instead of a linear program.

C. Post Processing

The post processing in [13] rounds up the continuous relaxation of user associations to binary associations. This is not an issue in the model considered here since user association over a slow timescale is indirectly determined by the amount of resources allocated to each group. However, a different post processing can be used to improve the delay performance without increasing the energy cost, since the objective (P1a) only depends on the binary variables \mathbf{z} .

Here post processing is executed after a solution of P1 is obtained. The process is performed over the subset of active BTSs, with all off BTSs removed from the formulation. Specifically, the post processing is to solve the following problem (cf. [9]):

$$\begin{aligned} & \underset{\mathbf{r}, \mathbf{x}, \mathbf{y}}{\text{minimize}} && \sum_{j \in \mathcal{K}} \frac{\lambda_j}{\sum_{l \in \mathcal{K}} \lambda_l} \cdot \frac{1}{r_j - \lambda_j} \\ & \text{subject to} && \text{(P1c)–(P1h)} \end{aligned} \quad (\text{P5a})$$

where objective (P5a) is the average packet sojourn time in the network. P5 is a convex program and is much easier to solve than P1.

³ \mathbf{z}^t corresponds to the (possibly reduced) set \mathcal{N}_2 . Here $f^1(\mathbf{z}^t)$ is evaluated by setting any z_i^t not in the original \mathcal{N}_2 to zero.

TABLE I
PARAMETER CONFIGURATIONS.

Parameter	Value/Function
macro transmit power	46 dBm
pico transmit power	30 dBm
total bandwidth	10 MHz
average packet length	0.5 Mb
macro to UE pathloss	$128.1 + 37.6 \log_{10}(R)$
pico to UE pathloss	$140.7 + 36.7 \log_{10}(R)$

We shall see in Section V that post processing can greatly improve the delay performance in the light traffic regime. The post processing is helpful if the objective only depends on the ℓ_0 norm. We can also consider the generalizations in Section III-D to minimize the total combined cost of energy and delay.

V. NUMERICAL RESULTS

In this section, we demonstrate the effectiveness of the proposed method through numerical examples. In particular, we compare the performance of Algorithms 1 and 2 for solving P1 as well as the performance of Algorithm 1 for solving P3, which basically corresponds to the scheme of [13].

A. Simulation Setup

The simulation is carried out over the network in Fig. 1. The HetNet is deployed on a $500 \times 1000 \text{ m}^2$ area. The entire geographic area is divided into a hexagonal grid with 66 hexagons. Each hexagon represents a user group. The UEs within each hexagon/user group are assumed to be at the center of the hexagon. The BTSs are assumed to be at the vertices of the hexagons. There are two macro BTSs in the HetNet denoted by the dark squares. Ten pico BTSs are randomly placed in the network denoted by the triangles.

The spectral efficiency is calculated by (2) with a 30 dB cap on the receive SINR (i.e., an SINR greater than 30 dB is regarded as 30 dB). The pathloss models used for macro and pico BTSs are the urban macro (UMa) and urban micro (UMi) models specified in [29]. Other parameters used in the simulation are given in Table I, which are also compliant with the LTE standard [29]. In the simulation, we only consider pathloss without slow and fast fading.

B. Energy-Efficient Spectrum Allocation

Fig. 2 illustrates the energy cost due to different allocation schemes at different traffic intensities. In the simulation, a random vector $[a_1, \dots, a_k]$ is first generated with $\mathbb{E}[a_i] = 1$, $i = 1, \dots, k$. Given the average arrival rate $\bar{\lambda}$, let $\lambda = \bar{\lambda}[a_1, \dots, a_k]$, i.e., the arrival rates of all user groups are scaled proportionally with $\bar{\lambda}$. The delay requirement for all UEs is 0.5 seconds, i.e., $\tau_j = 0.5$, $\forall j \in \mathcal{K}$. The energy costs of all pico BTSs are set at one unit, i.e., $c_i = 1$, $\forall i \in \mathcal{N}_2$. Hence the y axis also indicates the number of active pico BTSs. The curve marked by circles is obtained by solving P3 using Algorithm 1, which can be interpreted as applying the scheme of [13] to the scenarios presented in this paper. The two curves marked by squares and triangles are achieved by

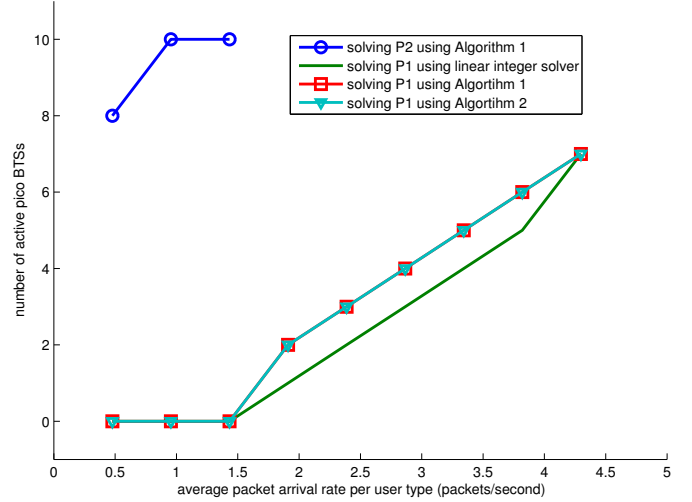


Fig. 2. Comparison of energy costs of different allocation schemes.

solving P1 using Algorithms 1 and 2, respectively. These two curves coincide because those two algorithms yield the same energy savings in this case. The curve without any marker is achieved by solving P1 using a standard integer program solver (to be specified shortly).

According to Fig. 2, the solution to P1 greatly outperforms the solution to P3. The solution to P3 can only support up to 1.4 packets/second per user group. In contrast, the proposed scheme here can serve up to 4.3 packets/second per user group, a three-fold throughput gain.

The results obtained using Algorithms 1 and 2 to solve P1 are very close to the solution obtained using a standard integer program solver. (We observe that at most one extra pico BTS is turned on in all traffic regimes.) Similar results are observed with random realizations of the network and traffic distribution.

The spectrum allocation and user association according to the solution to P1 are shown in Fig. 3. Each pie chart indicates the spectrum allocation at the corresponding user group. The different colors represent different patterns, and the amount of spectrum resources allocated to each group under each pattern is denoted by the size of the corresponding pie. The average packet arrival rate of each group is shown by the number above the pie chart. Each line segment joining a BTS and a group means the group is served by the BTS. The color bars on top of each figure shows the actual spectrum partition into different patterns.

In both Fig. 3a and Fig. 3b, the spectrum resources allocated to each user group is roughly proportional to the corresponding traffic demand. The light traffic scenario is shown in Fig. 3a. All the pico BTSs are turned off, leaving the 2 macro BTSs to serve all the user groups. The spectrum in Fig. 3a is divided into two segments each exclusively used by one of the macro BTSs. Apparently, the allocation is suboptimal in terms of user association, since both macro BTSs intrude into the other cell to serve some user groups that should obviously be served by the other macro BTS. This is because the objective in P1 is only to minimize the number of active pico BTSs. The

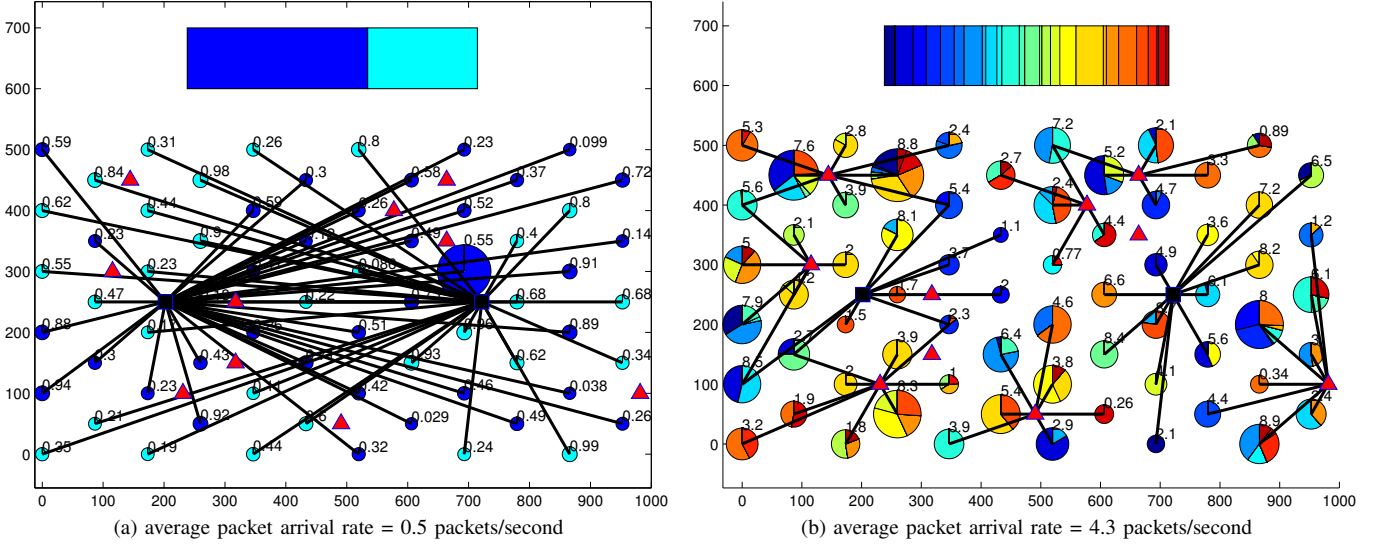


Fig. 3. Spectrum allocation and user association before post processing in light and heavy traffic regimes.

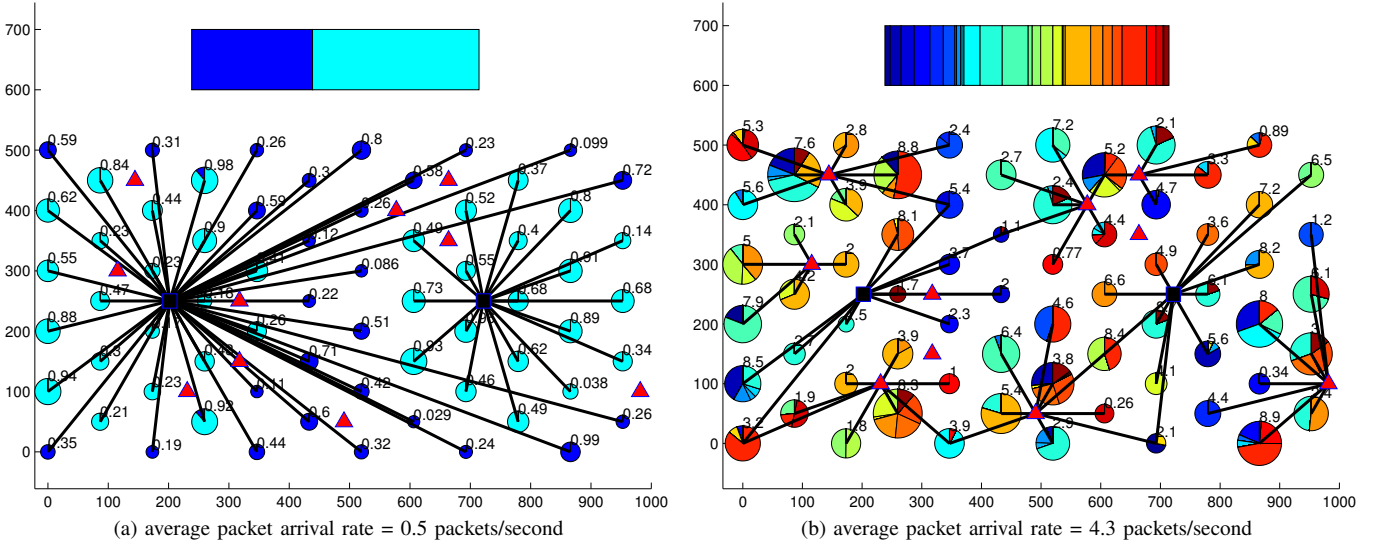


Fig. 4. Spectrum allocation and user association after post processing in light and heavy traffic regimes.

algorithm terminates as soon as a feasible spectrum allocation is found that satisfies the user delay requirements using the minimum number of pico BTSs. The heavy traffic scenario is shown in Fig. 3b. The spectrum allocation is topology aware. Each pico BTS serves nearby user groups. Macro BTSs then serve the user groups in coverage holes of the pico BTSs. The two pico BTSs turned off are near clusters of other pico BTSs. The spectrum is orthogonalized among nearby user groups, and is efficiently reused by user groups that are far away.

The spectrum allocations in the light and heavy traffic regimes after the post processing for further delay improvement, as described in Section IV-C, are shown in Fig. 4. The allocation in the light traffic regime is shown in Fig. 4a. The spectrum is still divided into 2 segments. However, instead of assigning each segment to each BTS exclusively, one segment is shared by both BTSs and the other is exclusively used by the macro BTS on the left. The post processing achieves

fractional frequency reuse, i.e., full frequency reuse among all user groups in cell centers, whereas user groups at cell edges are served with the spectrum exclusively allocated to the left macro BTS. The reason that cell edge users in the right macro cell are also served by the left macro BTS is due to the relatively small network size.

The spectrum allocation after post processing shown in Fig. 4b is similar to that in Fig. 3b. This is because as the traffic gets close to the maximum throughput, there is little margin to further reduce the delay once the delay requirements are met. The post processing reduces the average packet sojourn time in the network from 0.49 seconds to 0.29 seconds in the light traffic regime. However, it only reduces the delay from 0.5 seconds to 0.45 seconds in the heavy traffic regime.

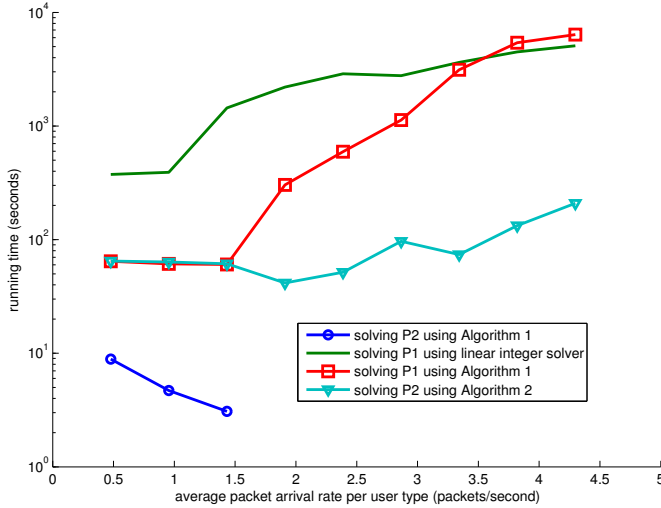


Fig. 5. Runtime comparison at different traffic intensities.

TABLE II
NUMBER OF ITERATIONS USING ALGORITHM 1 AND ALGORITHM 2.

arrival rate (packets/second)	0.48	1.43	2.39	3.34	4.30
iterations for Algorithm 1	2	2	13	57	102
iterations for Algorithm 2	2	2	9	6	13

C. Runtime Considerations

The runtimes of the four schemes are shown in Fig. 5. The curves are marked in the same way as in Fig. 2. The optimization problems were solved in Matlab using CVX from CVX Research, Inc. on an Intel Core i7 2.4 GHz quad-core computer with 16 GB RAM. The continuous linear program P4 in the iterations of Algorithms 1 and 2 was solved by the default linear program solver of Gurobi [30]. The parameters in Algorithms 1 and 2 are $T = 200$, $\epsilon_1 = \epsilon_2 = 10^{-9}$ and $\alpha = 0.1$. The curve without marker was solved by the default linear integer program solver of Gurobi.

Although the solution to P3 is suboptimal in both energy savings and network throughput, it can be solved much faster than the other three, and can be applied to large systems with hundreds of BTSs [13]. Among the remaining three algorithms, Algorithm 2's runtime is the most favorable under all traffic conditions.

The number of iterations when using Algorithms 1 and 2 is shown in Table II. In the light traffic regime, both algorithms converge within two iterations. In fact, both algorithms immediately realize that all pico BTSs should be turned off under such light loads. Hence the reduction of the feasible set in Algorithm 2 never happens. In the moderate and heavy traffic regimes, Algorithm 2 takes fewer iterations due to the feasible set reduction. The runtime using Algorithm 2 is reduced by as much as 42 times compared to Algorithm 1 as shown in Fig. 5. The shorter runtime is due to both faster convergence and lower computational complexity in each iteration with reduced dimensions.

D. Overhead

A central controller needs to know the traffic intensity of

all UE groups and the spectral efficiency of all links (i.e., λ and s) in order to perform the proposed energy-efficient global resource management. Location information can help to identify users from different groups, which can be acquired using standard positioning schemes. The traffic information for each UE group can be measured at its serving BTS. The link gains are routinely measured by the BTSs. If BTS i does not receive signals from UE group j , then the gain of link $i \rightarrow j$ can be regarded as zero.

On the slow timescale considered, the overhead of storing and forwarding the aforementioned information over backhaul links is quite small. For example, to describe 10,000 parameters (32 bits each) once every minute translates to about 5 kilobits per second (kbps). The decision variables need to be fed back to each BTS after solving the optimization problem. The number of variables sent to each BTS is at most $O(k^2)$. Even with a million variables per minute, the overhead is merely 500 kbps.

VI. CONCLUSION

Traffic-driven radio resource management on a slow timescale has been proposed and studied for improving energy efficiencies in HetNets. Joint spectrum allocation, user association, and cell activation significantly reduce energy cost and improve system throughput. The proposed algorithms can efficiently optimize a network cluster with up to 20 BTSs. The improved performance is at the cost of increased computational complexity for exploiting spectrum agility.

The main drawback of the current problem formulation is that the complexity scales exponentially with network size due to the combinatorial patterns. In practice, many of the 2^n patterns in a large HetNet can be easily ruled out, i.e., two BTSs far apart will not interfere with each other, and a UE will not be served by a BTS far away. Extending this approach to much larger HetNets is ongoing work. The general resource management framework has also been extended to mixed fast and slow timescales in preliminary work [31].

ACKNOWLEDGEMENT

The authors thank Ermin Wei for pointing out a flaw in an earlier proof of Theorem 1.

REFERENCES

- [1] A. Fehske, G. Fettweis, J. Malmudin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55–62, 2011.
- [2] L. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Góanddor, G. Auer, and L. Van Der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Communications Magazine*, vol. 48, pp. 66–72, November 2010.
- [3] G. Y. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," *IEEE Wireless Commun. Mag.*, vol. 18, no. 6, pp. 28–35, 2011.
- [4] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, 2011.
- [5] S. McLaughlin, P. M. Grant, J. S. Thompson, H. Haas, D. I. Laurenson, C. Khirallah, Y. Hou, and R. Wang, "Techniques for improving cellular radio base station energy efficiency," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 10–17, 2011.

- [6] R. Q. 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, 2014.
- [7] P. Frenger, P. Moberg, J. Malmudin, Y. Jading, and I. Godor, "Reducing energy consumption in LTE with cell DTX," in *IEEE 73rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2011.
- [8] D. Astely, E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, pp. 44–51, april 2009.
- [9] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun. Special Issue on Recent Advances in Heterogeneous Cellular Networks*, vol. 33, no. 10, pp. 2027–2038, 2015.
- [10] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. IEEE International Conference on Communications Workshops*, pp. 1–5, june 2009.
- [11] S. Zhou, J. Gong, Z. Yang, Z. Niu, P. Yang, and D. Corporation, "Green mobile access network with dynamic base station energy saving," *Proc. ACM MobiCom*, vol. 9, no. 262, pp. 10–12, 2009.
- [12] 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, 2011.
- [13] E. Pollakis, R. L. G. Cavalcante, and S. Stanczak, "Base station selection for energy efficient network operation with the majorization-minimization algorithm," in *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, pp. 219–223, IEEE, 2012.
- [14] Z. Niu, S. Zhou, Y. Hua, Q. Zhang, and D. Cao, "Energy-aware network planning for wireless cellular system with inter-cell cooperation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1412–1423, 2012.
- [15] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, 2013.
- [16] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, 2013.
- [17] Z. Yang and Z. Niu, "Energy saving in cellular networks by dynamic RS-bs association and bs switching," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4602–4614, 2013.
- [18] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour, "Traffic-aware two-dimensional dynamic network provisioning for energy-efficient cellular systems," *Transactions on Emerging Telecommunications Technologies*, 2014.
- [19] X. Li, X. Zhang, and W. Wang, "An energy-efficient cell planning strategy for heterogeneous network based on realistic traffic data," in *International Conference on Computing, Management and Telecommunications*, pp. 122–127, 2014.
- [20] G. Lim, C. Xiong, L. J. Cimini Jr., and G. Y. Li, "Energy-efficient resource allocation for OFDMA-based multi-RAT networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2696–2705, 2014.
- [21] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and resource allocation in heterogeneous networks via sparsity pursuit," <http://arxiv.org/abs/1408.5091>, 2014.
- [22] A. Dotzler, W. Utschick, and G. Dietl, "Fractional reuse partitioning for mimo networks," in *Proc. IEEE GLOBECOM*, pp. 1–5, Dec 2010.
- [23] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
- [24] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine learning*, vol. 85, no. 1-2, pp. 3–39, 2011.
- [25] B. Zhuang, D. Guo, and M. L. Honig, "Energy management of dense wireless heterogeneous networks over slow timescales," in *Proc. Allerton Conf. Commun., Control, & Computing*, pp. 26–32, Oct. 2012.
- [26] Z. Zhou, D. Guo, and M. L. Honig, "Allocation of licensed and unlicensed spectrum in heterogeneous networks," in *preprint*, 2015.
- [27] R. Nelson, *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*. Springer, 1995.
- [28] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.
- [29] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," V9.0.0, March 2010.
- [30] Gurobi Optimization, "Gurobi optimizer reference manual," 2015.
- [31] F. Teng and D. Guo, "Resource management in 5G: a tale of two timescales," in *Proc. Asilomar Conf. Signals, Systems, & Computers*, Pacific Grove, CA, USA, 2015.