QoS-Aware Energy Efficient Association and Resource Scheduling for HetNets

Taewoon Kim and J. Morris Chang, *Senior Member, IEEE*

***Abstract*—A heterogeneous network (HetNet) can actively utilize the spectrum reuse with low power consumption, and thus is promising for the next-generation cellular networks. However, there are some technical challenges to be overcome in order for HetNets to be practical, and we address the following two in this paper. One is how to formulate the association and resource scheduling problem in a way that an optimal solution can be found in a reasonable amount of time, and the other is how to accommodate varying users' demand. In order to minimize the power consumption and to satisfy varying users' QoS (Quality of service) requirements, we propose a low-complex, distributed association and resource allocation scheme. By taking a cost-based approach, we first separate a non-convex joint association and resource allocation problem into two subproblems. The channel allocation and base station assignment problem is then relaxed so that the problem becomes tractable. For the power allocation problem, we introduce a low-complex iterative algorithm by using the decomposition theory. The evaluation results show that the proposed solution can maintain the overall power consumption minimized while satisfying the QoS requirements.**

***Index Terms*—Heterogeneous networks, smallcell, association, resource scheduling, power consumption, quality of service.**

I. Introduction

A Recent report pointed out that in 2014 the amount of data traffic from mobile devices increased 69%, resulting in a monthly usage of 2.5 exabytes [1]. Such an explosive increase in mobile traffic has been led by a widespread usage of mobile handheld devices and bandwidth-hungry applications. The ever-increasing mobile traffic is unlikely to be saturated at least for the next five years; rather, the total amount of traffic generated by mobile devices is expected to be increased by tenfold between 2014 and 2019 [1].

On one hand, experts from academia and industry are seeking for ways of boosting up the communication technology to be prepared for the sharply-increasing traffic demand. On the other hand, the increased energy consumption is a matter of concern to the environment because of the greenhouse gas emissions and the increasing OPEX[1]. It is estimated that the ICT sector accounts for 2% of the total CO_2 emission, and among those from networks, to be specific, the mobile communication infrastructures will contribute more than 50% by 2020 [2]. In particular, the power consumption from BSs is expected to account for 60–80% of the total power usage from cellular networks [3] due to the rapid growth in both the number of BSs deployed and the aggregate mobile traffic.

Table I
ABBREVIATIONS AND ACRONYMS

CAPEX	CAPital EXpenditure
HetNet	Heterogeneous Network
ICT	Information and Communications Technology
LTE	Long-Term Evolution
MBS	Macro Base Station
OFDMA	Orthogonal Frequency-Division Multiple Access
OPEX	OPerational EXpenditure
QoS	Quality of Service
SBS	Smallcell Base Station
SINR	Signal-to-Interference-plus-Noise Ratio
UE	User Equipment

Among those technologies taking the two aforementioned aspects into account, the heterogeneous architecture [4] is one of the most promising technologies, because it is not only applicable to the current 4G networking system, but also considered as an essential component for the future generation (5G) networking system [5] [6] [7]. A HetNet, in general, is formed by deploying multiple low-power, low-cost SBS (e.g., microcells, picocells and femtocells) on top of a high-power MBS, and it has many advantages. For example, due to the short coverage of SBSs, the spectrum reuse can be actively exercised. Also the channel quality between an SBS and its associated UE is so good that a higher data rate can be easily achieved, while operating in low power. Being relatively compact in size as well as the ease of installation enables SBSs to be flexibly deployed so that they can effectively extend the coverage of MBSs and offload users' traffic especially in crowded areas, such as shopping malls, sport stadiums, concert arenas and so forth. Further, much cheaper CAPEX and OPEX of SBSs have intrigued a great amount of attention from both academia and industry.

Motivated by the fact that HetNet is a cost-effective and practical solution, the use of SBSs has been widely introduced (including [5], [7] and [8]) and studied in many literatures (please refer to Section II. Related Work). In addition, the concept of HetNet is accepted by the standard body and introduced in LTE [4] [9] [10]. One common concern in those works is either how to offload the user traffic from an MBS to SBSs or how to schedule the network resource so as to achieve their own objectives, such as maximizing spectral efficiency, energy efficiency, a certain utility measure and so forth.

As mentioned in [11], however, introducing SBSs may rather increase the overall power consumption unless they are handled carefully. In general, the careful handling includes well-designed offloading strategies (i.e., establishing associations between UEs and SBSs), preferably with a dynamic

T. Kim is with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA. E-mail: tkim@iastate.edu.

J. M. Chang is with the Department of Electrical Engineering, University of South Florida, Tampa, FL 33647, USA. E-mail: chang5@usf.edu.

[1] Abbreviations and acronyms are listed in Table I.

switching on/off scheme [2] [3], and resource scheduling (i.e., transmission power control and bandwidth allocation). Even after the optimal decision has been made, one cannot guarantee the optimal decision will remain optimal in the future because the QoS requirement of each user and/or the wireless link quality frequently changes over time. However, a frequent manual reconfiguration of the system is not appealing because it is hard to be responsive to the network dynamics, and also increases OPEX to a large extent especially when the BSs are densely deployed [10]. In this regard, a self-configuring or an automated mechanism that is able to respond or adapt to the network dynamics needs to be studied from the perspective of HetNets.

In addition, the centrality can cause a serious issue especially in HetNets. In a centralized system, all decisions are made by one or a small group of entities. As a result, a huge volume of information should be either exchanged in real-time or stored/updated at a shared storage, which incurs a significant burden to the system. In addition, a centralized system requires a high processing capability and power consumption to handle a large volume of data as the network grows in size. For those reasons, a centralized system may not be responsive and it might even consume a large amount of power for resource scheduling; whereas a distributed algorithm does not.

In this paper, we study the distributed user association and resource allocation for HetNets that minimizes the overall power consumption. By considering both association and resource scheduling together, we propose a complete management framework for an energy-efficient HetNet. In order to efficiently schedule the use of BSs as well as the networking resource (i.e., transmission power and spectrum) we formulate a two-stage iterative optimization problem which can be solved in a distributed manner without requiring a heavy control message exchange or a high computational cost. To do so, we first partition the non-convex, joint association and resource allocation problem into two subproblems by using a cost-based approach that effectively estimates the power use. In addition, relaxation and decomposition techniques are applied to the association and the resource scheduling problems, respectively, so as to reduce the computational complexity. After the decomposition, we propose a distributed power update method, which converges to the optimum; we show its convergence by both simulation and analysis. An extensive amount of evaluations and comparison studies has been performed on various network scenarios to show the effectiveness of the proposed BS/channel assignment and power allocation scheme. Lastly, we show the efficiency of the proposed scheme by showing its fast convergence.

The rest of this paper is organized as follows. Section II introduces some of the relevant literatures to the proposed method in this paper. In Section III, we describe the network model and then introduce the problem formulation for both i) BS association and channel assignment and ii) power allocation. Section IV presents the evaluation and comparison results, and finally Section V concludes the paper.

II. Related Work

In addition to the research on the energy-efficient channel assignment and power allocation for homogeneous cellular networks [12] [13], a large volume of studies have been done on two-/multi-tier HetNets aiming at optimizing the user association (referred to as access control) and/or resource scheduling (referred to as resource allocation). Zhang et al. [14] proposed a user association and interference management scheme that maximizes the sum utility of the average achievable rates. A group muting scheme is also used to reduce the interference between nearby SBSs, but the proposed method does not consider the QoS requirements. The authors in [15] proposed a game-theoretic approach to formulate a user association problem that maximizes the throughput. They further showed the practicality of the proposed method by studying its convergence behavior, yet leaving the QoS requirements unexplored. Singh et al. [16] proposed a general model for the joint resource partitioning and offloading on a two-tier HetNet. They derived the optimal strategy that improves the rate of cell edge users. However, their work did not consider the QoS requirements.

The work in [17] proposed a cell association and resource allocation scheme for downlink HetNets that balances the network traffic. They also developed a distributed algorithm, yet leaving the QoS requirements unexplored. Femto-matching [18] is an auction-based algorithm for load balancing and fair resource sharing among BSs and UEs, respectively, on HetNets. The authors also proposed a polynomial-time solution by transforming the initial problem formulation. However, the evaluation has been done on a simple network where there is only a single MBS, and QoS requirements are not considered. Shen et al. [19] studied the joint association and power control problem with beamforming, and then proposed an iterative method for the corresponding problem. The problem formulation therein maximizes the network utility function considering the proportional fairness. However, the QoS constraints are left unexplored.

The work in [20] considers a joint channel allocation and power control problem for femtocell networks. The proposed solution maximizes the total minimum spectral efficiency, and the corresponding distributed algorithm is also developed. However, the work did not pay attention to satisfying QoS requirements in particular for the femtocell UEs. Ngo et al. [21] proposed a joint subchannel and power allocation scheme for downlink HetNets. The proposed algorithm maximizes the total throughput for the second-tier UEs, while causing no performance degradation to the first-tier UEs. The limitation of the work is twofold. One is that it assumes the fixed power allocation, and the other is that it does not consider the QoS requirements for the second-tier UEs.

The proposed work in [22] formulated the throughput maximization problem, subject to QoS requirements for two-tier femtocell networks. It also studied the effect of different femtocell access policies (i.e., open and closed) on the overall throughput performance. However, the work in [22] may encounter a scalability issue since it lacks a distributed mechanism which is very important, in particular, when the network size becomes large.

Bao et al. [23] proposed an optimal resource allocation scheme on HetNets that maximizes the downlink sum throughput. The authors also considered both spatial and temporal dimensions, i.e., random distribution of BSs and dynamic user traffic session arrivals in time, respectively. However, the absence of a distributed counterpart may cause a scalability issue on a large-scale network. The authors in [24] proposed a framework on which they studied the joint association and resource allocation problem for HetNets. By making use of the framework, they compared the different channel allocation strategies and different association rules with each other. The proposed work therein is centralized, which may not be scalable with respect to the network size.

Chandrasekhar et al. [25] proposed an optimal spectrum allocation policy for HetNets that maximizes the area spectral efficiency with the QoS requirement considered. However, their work is limited in that SBSs use a simple channel allocation policy, i.e., Round Robin, and all channels are assigned equal transmission powers which may not achieve the optimal performance. Zhuang et al. [26] proposed a traffic-adaptive resource allocation algorithm that schedules the spectrum resource in an adaptive manner, subject to the network layer QoS, i.e., delay. A simple power allocation scheme is used such that it assigns an equal power over the spectrum, which may not achieve the optimal performance. Abdelnasser et al. [27] proposed a power and channel allocation scheme for a two-tier HetNet. They formulated a tier-aware resource allocation problem subject to QoS requirements, and then proposed a distributed algorithm. However, their network model is not practical in that they consider only a single MBS, and they also assume equal power allocation for MBS UEs which may not achieve the optimal performance.

Y. Li et al. [28] studied a QoS-guaranteed D2D (device-to-device) network underlying a cellular network. The proposed joint admission control and resource allocation problem in [28] is decoupled into four subproblems, i.e., mode selection, admission control, partner assignment and power allocation, to make the problem tractable. In addition, they proposed a fast heuristic algorithm to further reduce the computational cost. The authors considered a uplink transmission in a single-tier D2D-enabled network, whereas we focus on two-tier downlink cellular networks in this paper. We also consider the general network configuration where there are multiple BSs are deployed, while the work in [28] considered a single BS scenario. Son et al. [29] studied an interference management scheme for downlink heterogeneous networks, and proposed REFIM (REFerence based Interference Management). The proposed user scheduling and power allocation method in [29] is converted to a low-complex algorithm by using the notion of reference users. REFIM is a weighted throughput maximization problem, and does not consider the QoS requirements. Also, REFIM considers only a single user with the maximum channel gain when allocating transmission power for each channel, while the proposed method in this paper considers the actual amount of interference in order to satisfy QoS requirements and to minimize the power consumption. Li et al. [30] formulated a stochastic optimization problem for energy-efficient operations for HetNets considering both spatial and temporal traffic fluctuations. The proposed algorithm, SEED (Steerable Energy ExpenDiture), decouples optimization variables to reduce the computational complexity of the user association and subcarrier assignment subproblems. In addition, a sequential approximation and a greedy heuristic approach are used for power allocation and BS operation problems, respectively. Although the the proposed scheme in [30] tries to stabilize the network by assuring a finite average delay, it does not take the QoS requirements of individual users into account. For other previous works that are not discussed here, please refer to [27].

Although the aforementioned previous literatures studied two-tier HetNets from various perspectives, the proposed work in this paper has made the following advancement. We consider both user association and resource allocation together, and the proposed method allows UEs to be dynamically offloaded to SBSs, which all together is expected to enhance the capacity of the cellular networks to a large extent. Our work directly focuses on the energy minimization which has gained more and more attention recently, and also we consider many practical constraints. The proposed scheme considers the users' QoS requirements which has been ignored in many literatures. In addition, it allocates optimal power levels to different UEs instead of assigning equal powers to all channels. Thus, the proposed method in this paper is expected to fulfill the service requirements of users, while spending no more than necessary amount of power. In addition, a novel lightweight, distributed mechanism is provided with the proof of convergence, all of which are crucial as the network grows in size. The proposed scheme is evaluated under various and practical scenarios with a realistic channel model. Lastly, through comprehensive performance comparison, we show that the proposed scheme can effectively schedule the networking resource on HetNets.

III. Problem Formulation

We begin this section by describing the network model and assumptions. In what follows, we introduce how the optimal user association and resource allocation problem are formulated.

A. Network Model and Assumptions

Throughout the paper, we focus on the downlink transmission for a two-tier OFDMA (Orthogonal Frequency-Division Multiple Access) cellular network. On the network are M MBSs, and each of which is overlaid by S SBSs. MBSs are distributed in a planned manner (e.g., by keeping the same inter-cell distance between the nearby MBSs) in order to provide area coverage, mobility management and so forth; while SBSs are randomly[2] distributed [4] [14] [17] [18] [31] [32] [33]. The reason for assuming the random deployment of SBSs is that their deployment is much less planned [11] compared to that of MBSs. To be specific, they are likely to be installed on demand or in an ad hoc manner so as to

[2] In Section IV, we have used Uniform distribution for simulation.

fulfill a sudden or periodic increase of the QoS requirement in certain areas such as a shopping center, sports complex, office, household and so forth. It is worth mentioning that the proposed scheme does not assume or rely on any specific area or region. Therefore, in order to show that the proposed scheme does not depend on any certain distributions of SBS, a random distribution is used to represent the placement/layout of SBSs in general. All MBSs are always active in order to provide the full area coverage to all UEs [10]. On the other hand, SBSs may or may not be active depending on the user association status at the moment. Any non-offloaded UEs are automatically associated with the MBS that provides the strongest signal strength by default.

MBSs and SBSs are assumed to operate on different frequency bands to avoid cross-tier interference [22] [23] [25] [33]. However, the same type of base stations share the same frequency range, and thus they always interfere with each other. In this work, the *coverage* of an MBS indicates the area within which all non-offloaded UEs shall associate with the MBS. However, the signal generated by each MBS propagates beyond its coverage, and thus incurs interference to the rest MBSs; this principle applies to SBSs as well. Both types of base stations can access the core network through wired communication links. Each MBS has U UEs (or users) whose average data rate requirements are known. We assume that SBSs operate fully (or in part) with an open access mode.[3] The available bandwidth is divided into multiple channels, each of which is Δf-wide in Hz. We also assume the continuous power and rate control. Some of the frequently-used notations are summarized in Table II, and other notations that are not on the table will be introduced when necessary.

Table II
SUMMARY OF NOTATIONS

M	Number of MBSs on the network
S	Number of SBSs on a macrocell
U	Number of UEs on a macrocell
N_{ch}	Number of available channels
N_{ch}^{M}	Number of available channels for an MBS
N_{ch}^{S}	Number of available channels for an SBS
$\mathcal{N}^{M}$	Index set of MBS-accessible channels
$\mathcal{N}^{S}$	Index set of SBS-accessible channels
$\mathcal{M}$	Index set of MBSs, $\{1, 2, \cdots, m, \cdots, M\}$
$\mathcal{S}^{m}$	Index set of SBSs overlaid on MBS m, $\{1, 2, ..., s, ..., S\}$
$\mathcal{U}^{m}$	Index set of UEs within the coverage of MBS m, $\{1, 2, ..., u, ..., U\}$
$\mathcal{U}_{0}^{m}$	Subset of $\mathcal{U}^{m}$. UEs associated with MBS m
$\mathcal{U}_{s}^{m}$	Subset of $\mathcal{U}^{m}$. UEs associated with SBS $s \in \mathcal{S}^{m}$
$\mathbf{p}_{0}^{m}$	Transmission power vector of MBS m over channels
$\mathbf{p}_{s}^{m}$	Transmission power vector of SBS $s \in \mathcal{S}^{m}$ over channels
$\mathbf{g}_{u}^{m}$	Channel gain vector of UE $u \in \mathcal{U}^{m}$ over channels
η_{thr}	SINR threshold
r_{u}^{m}	QoS (i.e., data rate) requirement of UE $u \in \mathcal{U}^{m}$
Δf	Channel bandwidth
P_{max}^{M}	Maximum transmission power of MBS
P_{max}^{S}	Maximum transmission power of SBS
$\mathbf{c}_{u}^{m}$	Cost vector of UE $u \in \mathcal{U}^{m}$ over channels
N_{0}	Per-Hz noise power

B. Cost-Based Problem Separation

Considering that the joint association and resource allocation problem belongs to a mixed integer nonlinear program with the decision variables coupled, the computational cost of the joint problem is prohibitive. Thus, the approach taken in this work is to first partition the problem into two, one for the user association and channel allocation (Stage 1), and the other for the power allocation (Stage 2), and then apply relaxation and decomposition techniques, respectively, in order to make the whole procedure tractable and suitable for online resource scheduling. To this end, we have introduced a cost function, by which the original problem will be separated into two. It is worth mentioning that after the partitioning, the proposed two-staged method may result in a sub-optimal solution. After the problem separation, however, Stage 1 does not know how much power will actually be used for communication. Therefore, it is crucial that the cost needs to be designed in a way that it can correctly estimate the amount of power to be used.

At the beginning of Stage 1, a UE $u \in \mathcal{U}^m$ senses the pilot signals from nearby BSs over the entire channels, and produces a channel gain vector $\mathbf{g}_u^m \in \mathbb{R}_{+}^{N_{ch}}$. Out of N_{ch} entries in $\mathbf{g}_u^m$, N_{ch}^M elements correspond to the measured channel gains between UE $u \in \mathcal{U}^m$ and MBS m over N_{ch}^M channels that are accessible to MBSs. Therefore, all the N_{ch}^M elements in $\mathbf{g}_u^m$ are strictly positive due to the area coverage provided by the closest MBS m, while the rest elements are nonnegative. If a UE resides within the coverage of an SBS, we have $\mathbf{g}_u^m \succ \mathbf{0}_{N_{ch}}$ where $\succ$ is an element-wise greater than operator and $\mathbf{0}_{N_{ch}}$ is an N_{ch}-by-1 zero vector. Also, in such cases, a UE can recognize the identifier of the SBS by decoding the pilot signal.

To be consistent throughout the paper, let us assume that the indices of the channels used by MBSs come before the ones used by SBSs. In other words, out of N_{ch} number of channels available whose index starts from 1 to N_{ch}, each MBS has an access to the channels indexed by $1, 2, \cdots, N_{ch}^M$, while an SBS is allowed to use the ones indexed by $N_{ch}^M + 1, \cdots, N_{ch}$. In this regard, let $\mathcal{N}^M$ and $\mathcal{N}^S$ be $\{1, 2, \cdots, n, \cdots, N_{ch}^M\}$ and $\{N_{ch}^M + 1, \cdots, n, \cdots, N_{ch}\}$, respectively.

Given $\mathbf{g}_u^m$ and the data rate requirement $r_u^m \in \mathbb{R}_{++}$, each UE u builds its own cost vector $\mathbf{c}_u^m \in \mathbb{R}_{++}^{N_{ch}}$, where its n-th entry is:

$$c_{u,n}^{m} = \begin{cases} (2^{\tilde{r}_u^m} - 1)/g_{u,n}^{m}, & \text{if } g_{u,n}^{m} > 0. \\ \infty, & \text{otherwise,} \end{cases} \tag{1}$$

where $\tilde{r}_u^m = r_u^m / \Delta f$ is a normalized data rate requirement. In fact, $\mathbf{c}_u^m$ is a measure of power required to satisfy the data rate requirement of UE $u \in \mathcal{U}^m$ across all channels with the interference and noise term ignored. According to the Shannon's well-known channel capacity formula, an achievable bit rate over a channel is defined as $C = B \log_2(1 + \frac{gP}{I+N})$, where C is channel capacity measured in bits per second (bps), g is channel gain, P is transmission power, I is interference and N is noise. In order not to violate the QoS

[3] If all the SBSs are deployed by the end-users, the open access mode may not be a practical assumption to make. However, by focusing on the scenario where all SBSs are deployed by the network operator, or considering only those SBSs operating with the open access mode (or the hybrid mode [11]), we argue that this assumption still holds.

requirement for each UE, we need to satisfy the constraint $r_n^m \geq C$. Since the minimum power is achieved when the QoS requirement is satisfied with equality, we can rewrite the Shannon's capacity formula as follows after replacing B and g with Δf and $g_{u,n}^m$, respectively, to be consistent in notation: $r_u^m = \Delta f \cdot \log_2(1 + \frac{g_{u,n}^m P}{I+N})$ or $\tilde{r}_u^m = \frac{r_u^m}{\Delta f} = \log_2(1 + \frac{g_{u,n}^m P}{I+N})$. After rearranging the equation, the minimum transmission power that satisfies the QoS requirement can be found by $P = (2^{\tilde{r}_u^m})(I+N)/g_{u,n}^m$. By assuming the sum of interference and noise is not dominant in determining the transmission power, we get the following relation which leads to how we defined the cost term in Eq. (1), i.e., $P \propto 2^{\tilde{r}_u^m}/g_{u,n}^m = c_{u,n}^m$.

After gathering the cost vectors from all UE ($\forall u \in \mathcal{U}^m$) along with their nearby SBS IDs, if applicable, an MBS m runs both Stage 1 and 2 in sequence which will be discussed as follows.

C. Stage 1: User Association and Channel Assignment

The goal of this stage is to find the best association (i.e., an offloading strategy) and channel assignment that minimizes the overall cost, which represents the expected amount of power use as discussed before. Given the cost vectors collected, we have the following optimization problem P. 2 for each MBS m that minimizes the overall cost of making user association and channel assignment. Please note that *s.t.* in the problem formulation stands for *subject to*.

$$\min_{\mathbf{X}^m} \quad tr[\mathbf{X}^m \cdot (\mathbf{c}^m)^T] \tag{2a}$$

s.t.

$$\sum_{n=1}^{N_{ch}} X_{u,n}^m = 1, \forall u \in \mathcal{U}^m \tag{2b}$$

$$\sum_{u \in \mathcal{U}^m} X_{u,n}^m \leq 1, \forall n \in \mathcal{N}^M \tag{2c}$$

$$\sum_{u \in \mathcal{U}^m} X_{u,n}^m \cdot I_{u,s}^m \leq 1, \forall n \in \mathcal{N}^S, \forall s \in \mathcal{S}^m \tag{2d}$$

$$\mathbf{X}^m \in \{0,1\}^{U \times N_{ch}}, \tag{2e}$$

where $tr[\cdot]$ is the trace function that sums the diagonal elements of a matrix $\cdot$, the decision variable $\mathbf{X}^m$ is a U-by-N_{ch} matrix of which (u,n) element is 1 (or 0) if UE u is (or is not) assigned to channel n, $\mathbf{c}^m$ is a U-by-N_{ch} matrix whose u-th row corresponds to the cost vector of UE u and $\mathbf{I}^m$ is an U-by-S matrix whose (u,s) element is 1 if UE u successfully decoded the pilot signal from SBS s. The objective function (2a) in P. 2 calculates the total cost with respect to the mapping between UEs and channels (and BS as well). The objective function can also be written as $\sum_{\forall u \in \mathcal{U}^m} \mathbf{X}_u^m \cdot (\mathbf{c}_u^m)^T$ or $\sum_{\forall u \in \mathcal{U}^m} \sum_{n=1}^{N_{ch}} X_{u,n}^m \cdot c_{u,n}^m$. Since the cost $X_{u,n}^m \cdot c_{u',n'}^m$ is meaningful only when $u = u'$ and $n = n'$, we take the sum of only the diagonal elements from $\mathbf{X}^m \cdot (\mathbf{c}^m)^T$, i.e., $tr[\mathbf{X}^m \cdot (\mathbf{c}^m)^T]$. Each UE is allowed to use 1 unit of channel resource (2b), and each MBS and SBS channel cannot be used for more than 1 unit each, (2c) and (2d), respectively. The decision variable represents a membership relation, and thus is binary (2e).

Please note that P. 2 runs on a small time scale; for example, 1 ms to comply with the 3GPP E-UTRA requirement [9]. Given that 3GPP E-UTRA makes use of physical resource blocks for communication, even when the number of available channels is less than that of active UEs, the proposed method can still fulfill the service demand from UEs by scheduling the resource blocks. As long as the demand from a UE can be satisfied without violating the delay constraint, the proposed method may schedule the UE for communication in one of the following time slots if the number of available channels at the moment is not enough. That is, having failed in assigning a BS/channel to a UE for the moment does not necessarily mean a failure in satisfying the UE's QoS requirement. In addition, the densely deployed, short-range SBSs can achieve high frequency reuse, meaning that the aggregate number of channels seen by users can be larger than that of physical channels. However, if the aggregate service demand for a certain period exceeds the maximum attainable throughput over the network during the same period, some of the active users may experience service degradation, which will be discussed in Section III-D.

Due to the combinatorial nature of P. 2, however, the problem is not tractable, and thus is not suitable for an online scheduling. By relaxing the binary constraint (2e), we get the following convex problem that can be efficiently solved by each MBS m with the complexity of $\mathcal{O}((U \cdot N_{ch})^3)$ when the interior point method is used.

$$\min_{\mathbf{X}^m} \quad tr[\mathbf{X}^m \cdot (\mathbf{c}^m)^T] \tag{3a}$$

s.t.

$$\sum_{n=1}^{N_{ch}} X_{u,n}^m = 1, \forall u \in \mathcal{U}^m \tag{3b}$$

$$\sum_{u \in \mathcal{U}^m} X_{u,n}^m \leq 1, \forall n \in \mathcal{N}^M \tag{3c}$$

$$\sum_{u \in \mathcal{U}^m} X_{u,n}^m \cdot I_{u,s}^m \leq 1, \forall n \in \mathcal{N}^S, \forall s \in \mathcal{S}^m \tag{3d}$$

$$\mathbf{X}^m \in [0,1]^{U \times N_{ch}}. \tag{3e}$$

Although the optimal solutions from both P. 2 and P. 3 indicate the channel assignment as well as the user association for each UE, both solutions are not the same in practice. The binary solution from P. 2 lets each UE use the assigned channel and BS for a unit time, whereas the (possibly) non-binary solution from P. 3 forces a UE to hop between channels (and possibly between BSs as well) during the same unit time since the optimal solution indicates the fraction of time that a UE is allowed to use one or more BSs and channels. The non-binary solution seems to be attractive since it yields a better (or at least the same) optimal value because of the relaxation. However, it increases the amount of control messages as well as the scheduling complexity due to the frequent handover and channel hopping that should be made on a very precise timescale.

In this regard, we will recover binary solutions from the non-binary solutions from the relaxed optimization problem P. 3 by using the one-by-one removal algorithm [27] [34] [35] [36]. This relax-and-recover approach will help the system maintain a low complexity in both computation and operation.

Algorithm 1 Gradual one-by-one removal (for MBS m)

```
1: repeat
2:     Solve the relaxed optimization problem P. 3
3:     for $\forall u \in \mathcal{U}^m$ do
4:         $n^* = \arg\min_{\forall n} X^m_{u,n}$ such that $X^m_{u,n} \neq 0$
5:         Set $X^m_{u,n^*} = 0$
6:     end for
7: until all $X^m_{u,n}$ are binary
```

The Algo. 1 iteratively solves the relaxed optimization problem (line 2), searches for the nonzero minimum value for each UE (line 4), and forces each to be zero (line 5). Each MBS concurrently runs the algorithm which terminates in less than or equal to N_{ch} number of iterations. Please note that the solution found by running Algo. 1 may yield a sub-optimal solution to P. 2; the optimality of the solution will be investigated in Section IV. Given the recovered binary user association and channel assignment decision, the following Stage 2 allocates the minimum power to each UE. Note that the channel assignment problem also finds the best BS match for each UE. Those SBSs with no UE associated shall change their state into the SLEEP mode in order to save energy, while the other SBSs stay in the ACTIVE mode [11].

D. Stage 2: Power Allocation

This stage allocates the minimum power to each UE by taking both SINR and QoS requirements into account. In this regard, we first formulate the centralized power allocation problem where one or a small number of central entities have to control the downlink power for all BSs. In what follows, the centralized problem is decomposed into multiple low-complex subproblems which are scalable and suitable for online scheduling.

Before introducing the Stage 2 problem formulation, let us extend the notation of the channel gain so that we can comprehensively represent the gain between all the entities including that do not even belong to the same macrocell. As a reminder, the channel gain $\mathbf{g}^m_u$ represents the channel gain between UE $u \in \mathcal{U}^m$ and either MBS m or SBS $s \in \mathcal{S}^m$, where all of them are within the coverage of MBS m. This is because $\mathbf{g}^m_u$ is determined by overhearing the pilot signals over channels; thus, it should be coupled with the nearest MBS and SBS (if applicable). Let $G^{0,m}_{u,n}$ be the gain over channel n between UE u and MBS m, where u does not need to be a member of $\mathcal{U}^m$. In the same manner, let $G^{s,m}_{u,n}$ be the gain over channel n between UE u and SBS s, where u does not need to be a member of $\mathcal{U}^m$, but s must be a member of MBS m, i.e., $s \in \mathcal{S}^m$. Thus, for any UE $u \in \mathcal{U}^m$, we have $\mathbf{g}^m_{u,n} = G^{0,m}_{u,n}$ for any $n \in \mathcal{N}^M$. On the other hand, we have $\mathbf{g}^m_{u,n} \neq G^{0,m'}_{u,n}$ for any $n \in \mathcal{N}^M \bigcup \mathcal{N}^S$ if $m \neq m'$.

Given the BS association and channel assignment made in Stage 1, we have the power allocation problem P. 4 for MBS m and all SBSs therein (i.e., $\forall s \in \mathcal{S}^m$) that minimizes the overall power consumption.

$$\min_{\mathbf{P}^m} \sum_{n \in \mathcal{N}^M} P^m_{0,n} + \sum_{s \in \mathcal{S}^m} \sum_{n \in \mathcal{N}^S} P^m_{s,n} \tag{4a}$$

s.t.

$$\eta^m_{u,n} \geq \eta_{thr} X^m_{u,n}, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}^m_0 \tag{4b}$$

$$\eta^m_{u,n} \geq \eta_{thr} X^m_{u,n}, \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}^m_s, \forall s \in \mathcal{S}^m \tag{4c}$$

$$\Delta f \log_2(1+\eta^m_{u,n}) \geq r^m_u X^m_{u,n}, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}^m_0 \tag{4d}$$

$$\Delta f \log_2(1+\eta^m_{u,n}) \geq r^m_u X^m_{u,n}, \quad \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}^m_s, \forall s \in \mathcal{S}^m \tag{4e}$$

$$0 \leq P^m_{0,n} \leq P^M_{max}, \forall n \in \mathcal{N}^M \tag{4f}$$

$$\sum_{n \in \mathcal{N}^M} P^m_{0,n} \leq P^M_{max} \tag{4g}$$

$$0 \leq P^m_{s,n} \leq P^S_{max}, \forall n \in \mathcal{N}^S, \forall s \in \mathcal{S}^m \tag{4h}$$

$$\sum_{n \in \mathcal{N}^S} P^m_{s,n} \leq P^S_{max}, \forall s \in \mathcal{S}^m, \tag{4i}$$

where $P^m_{0,n}$ is the power allocated by MBS m over channel $n \in \mathcal{N}^M$, $P^m_{s,n}$ is the power allocated by SBS $s \in \mathcal{S}^m$ over channel $n \in \mathcal{N}^S$, and $\eta^m_{u,n}$ in Eq. (4b) and Eq. (4c) is the measure of SINR defined in Eq. (5a) and Eq. (5b), respectively. If a UE $u \in \mathcal{U}^m$ is to associate with an MBS m (i.e., $u \in \mathcal{U}^m_0$) on a certain channel $n \in \mathcal{N}^M$, the interference that the UE u will experience is related to the transmission power allocated to the same channel by the other MBSs $m' \neq m$, which corresponds to Eq. (5a). On the other hand, if a UE is coupled with an SBS on a certain channel $n \in \mathcal{N}^S$, it will sense the interference caused by all the other SBSs on the network that have allocated transmission power to the same channel as in Eq. (5b). To be specific, in Eq. (5b) the first term in the denominator measures the interference from other SBSs in the same macrocell, whereas the second term measures the interference from all SBSs that do not belong to the same macrocell. Note that the amount of interference to an SBS-associated UE is not significant mainly due to the low transmission power of SBSs, and the penetration loss of walls. For each UE that is associated with an MBS or SBS, its SINR should be greater than or equal to the predefined threshold, η_{thr}, as in Eq. (4b) and Eq. (4c), respectively. The QoS requirements of UEs that are associated with an MBS or an SBS should be satisfied according to Eq. (4d) or Eq. (4e), respectively. The transmission power allocated to a certain channel cannot exceed the power budget of an MBS or an SBS as in Eq. (4f) or Eq. (4h), respectively. Finally, the aggregate transmission power of an MBS or an SBS cannot be larger than its power budget as denoted by Eq. (4g) or Eq. (4i), respectively.

It is worth mentioning that solving P. 4 for MBS m and all active SBSs therein is not independent of that of others since each MBS m and all SBSs therein need to know the inter-tier interference from the rest MBSs and SBSs, respectively. Therefore, a single or a set of computing resource has to solve the network-wide power allocation problem, making the centralized approach impractical for an online resource

$$\eta_{u,n}^{m} = \begin{cases} \dfrac{G_{u,n}^{0,m} P_{0,n}^{m}}{\sum_{m' \neq m \in \mathcal{M}} G_{u,n}^{0,m'} P_{0,n}^{m'} + \Delta f \cdot N_0} & \text{if } u \in \mathcal{U}_0^m. \quad (5a) \\ \dfrac{G_{u,n}^{s,m} P_{s,n}^{m}}{\sum_{s' \neq s \in \mathcal{S}^m} G_{u,n}^{s',m} P_{s',n}^{m} + \sum_{m' \neq m \in \mathcal{M}} \sum_{s' \in \mathcal{S}^{m'}} G_{u,n}^{s',m'} P_{s',n}^{m'} + \Delta f \cdot N_0} & \text{if } u \in \mathcal{U}_s^m. \quad (5b) \end{cases}$$

scheduling. In what follows, we transform the centralized power allocation problem P. 4 into low-complex subproblems such that each subproblem can be quickly solved in a distributed manner.

In order to build a distributed system we decompose the centralized problem P. 4 by using the decomposition theory [37], and then transform it into low-complex subproblems that can be independently solved by each BS. The power allocation problem P. 4 which is for both MBS m and all SBSs therein (i.e., $s \in \mathcal{S}^m$) already has two sets of easily-separable components. The first term in the objective function (4a) along with the following four constraints (4b), (4d), (4f) and (4g) forms the MBS power minimization problem which is independent of that for SBSs, i.e., the remaining parts of the problem. Therefore, we can form the power allocation problem only for MBS m as follows.

$$\min_{\mathbf{P}_0^m} \sum_{n \in \mathcal{N}^M} P_{0,n}^m \tag{6a}$$

s.t.

$$\eta_{u,n}^m \geq \eta_{thr} X_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m \tag{6b}$$

$$\Delta f \log_2(1 + \eta_{u,n}^m) \geq r_u^m X_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m \tag{6c}$$

$$0 \leq P_{0,n}^m \leq P_{max}^M, \forall n \in \mathcal{N}^M \tag{6d}$$

$$\sum_{n \in \mathcal{N}^M} P_{0,n}^m \leq P_{max}^M. \tag{6e}$$

What is left in P. 4 after taking P. 6 out is the power minimization problem for all SBSs $s \in \mathcal{S}^m$, where its objective is to minimize the sum of transmission power used by all SBSs in macrocell m with the following constraints, (4c), (4e), (4h) and (4i). Minimizing the total power usage is equivalent to minimizing each individually. Also, the set of constraints for each SBS s is independent of that for the rest SBSs provided the transmission power of other SBSs are fixed. As a result, we have the power allocation problem for each SBS $s \in \mathcal{S}^m$ as follows which can be solved if the transmission power and the channel gain information of other SBSs are assumed to be known.[4]

$$\min_{\mathbf{P}_s^m} \sum_{n \in \mathcal{N}^S} P_{s,n}^m \tag{7a}$$

s.t.

$$\eta_{u,n}^m \geq \eta_{thr} X_{u,n}^m, \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}_s^m \tag{7b}$$

$$\Delta f \log_2(1 + \eta_{u,n}^m) \geq r_u^m X_{u,n}^m, \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}_s^m \tag{7c}$$

$$0 \leq P_{s,n}^m \leq P_{max}^S, \forall n \in \mathcal{N}^S \tag{7d}$$

$$\sum_{n \in \mathcal{N}^S} P_{s,n}^m \leq P_{max}^S. \tag{7e}$$

Although both P. 6 and P. 7 as they are cannot be further decomposed due to the coupling constraints in (6e) and (7e), respectively, we can use the decomposability structure by forming a Lagrangian of each to make both problems be decomposed. By relaxing (6e), the Lagrangian of P. 6 is given as below.

$$\min_{\mathbf{P}_0^m} \sum_{n \in \mathcal{N}^M} P_{0,n}^m + \lambda\Big(\sum_{n \in \mathcal{N}^M} P_{0,n}^m - P_{max}^M\Big) \tag{8a}$$

$$\text{s.t.} \quad \text{constraints in: } (6b), (6c), (6d),$$

where λ is a nonnegative Lagrangian multiplier. As a result, we have a Lagrange dual problem as follows.

$$\max_{\lambda \geq 0} \min_{\mathbf{P}_0^m} \sum_{n \in \mathcal{N}^M} P_{0,n}^m + \lambda\Big(\sum_{n \in \mathcal{N}^M} P_{0,n}^m - P_{max}^M\Big) \tag{9a}$$

$$\text{s.t.} \quad \text{constraints in: } (6b), (6c), (6d).$$

We assume that each BS has multiple processors or at least a single processor with the multithreading capability, each of which is dedicated to each channel for power update. The dedicated processor or thread to each channel is called *channel manager*, which is in charge of controlling the downlink transmission power of the assigned channel. At the lower level, the channel manager for channel $n \in \mathcal{N}^M$ solves the following power minimization problem if there is a UE u associated with the channel (i.e., $X_{u,n}^m = 1$).

$$\min_{P_{0,n}^m} \quad h_{0,n}^m(\lambda) = (1 + \lambda) P_{0,n}^m \tag{10a}$$

$$\text{s.t.} \quad \eta_{u,n}^m \geq \eta_{thr} \tag{10b}$$

$$\Delta f \log_2(1 + \eta_{u,n}^m) \geq r_u^m \tag{10c}$$

$$0 \leq P_{0,n}^m \leq P_{max}^M. \tag{10d}$$

Then, the higher level problem forms a maximization problem over the Lagrange multiplier as follows.

$$\max_{\lambda \geq 0} \quad h_0^m(\lambda) = \sum_{n \in \mathcal{N}^M} h_{0,n}^m(\lambda) - \lambda P_{max}^M \tag{11a}$$

Since the dual function $h_0^m(\lambda)$ is differentiable, the higher level problem can be solved with a gradient method of which update method is given below.

$$\lambda^{t+1} = [\lambda^t + \alpha^t(\sum_{n \in \mathcal{N}^M} P_{0,n}^{m*} - P_{max}^M)]^+, \tag{12}$$

where t is a nonnegative, integer-valued iteration count, α is a positive stepsize, and $[\cdot]^+ = \max\{0, \cdot\}$ is a projection operator to the nonnegative orthant. The initial λ can be set to some

[4] Please note that the proposed method does not directly solve P. 7 and thus, it does not really make such assumptions. In fact, P. 7 is one of the steps that we make to design the distributed power allocation method which will yield the global optimal solution without making the assumptions.

non-negative value, e.g., zero, and α can be a sufficiently small positive number; please refer to [37] for further details on the step-size. Then, the dual variable λ^t will converge to the dual optimal λ^* as $t \to \infty$ [37].

By taking a closer look at the lower-level problem P. 10, we can further simplify the power allocation procedure, and find a simple method to solve it by an even more efficient way than the decomposed ones. Since λ is nonnegative and common to all lower level problems, dropping the $1+\lambda$ term from the objective function does not change the optimal decision value. For a power minimization problem with an SINR constraint, the optimality is achieved when the constraint is satisfied with equality, which is also true for the same problem with a QoS constraint. Therefore, the solution of P. 10 can simply be found by:

$$P_{0,n}^{m*} = \min\{\max\{P_{0,n}^{m(s)}, P_{0,n}^{m(q)}\}, P_{max}^{M}\}, \tag{13}$$

where $P_{0,n}^{m(s)}$ and $P_{0,n}^{m(q)}$ are the solutions that satisfy SINR and QoS requirements, respectively, with equality.

In the same manner, we can derive the distributed power allocation method for each SBS. By relaxing (7e) which is the coupling constraint in P. 7, we have the following Lagrangian.

$$\min_{\mathbf{P}_s^m} \sum_{n\in\mathcal{N}^S} P_{s,n}^m + \lambda(\sum_{n\in\mathcal{N}^S} P_{s,n}^m - P_{max}^S) \tag{14a}$$
$$\text{s.t.} \quad \text{constraints in: } (7b), (7c), (7d),$$

where λ is a nonnegative Lagrangian multiplier. At the lower level, the channel manager for channel $n \in \mathcal{N}^S$ solves the following power minimization problem if there is a UE associated with the channel (i.e., $X_{u,n}^m = 1$).

$$\min_{P_{s,n}^m} \quad h_{s,n}^m(\lambda) = (1+\lambda)P_{s,n}^m \tag{15a}$$
$$\text{s.t.}$$
$$\eta_{u,n}^m \geq \eta_{thr} \tag{15b}$$
$$\Delta f \log_2(1+\eta_{u,n}^m) \geq r_u^m \tag{15c}$$
$$0 \leq P_{s,n}^m \leq P_{max}^S. \tag{15d}$$

Considering the optimality condition of the given power minimization problem P. 15, its solution can be found by the following simple method as we did for each MBS channel manager.

$$P_{s,n}^{m*} = \min\{\max\{P_{s,n}^{m(s)}, P_{s,n}^{m(q)}\}, P_{max}^{S}\}, \tag{16}$$

where $P_{s,n}^{m(s)}$ and $P_{s,n}^{m(q)}$ are the solutions that satisfy SINR and QoS requirements, respectively, with equality.

Although each channel manager considers and guarantees the per-channel power budget constraint (i.e., the maximum transmission power for the channel should not be greater than P_{max}^M or P_{max}^S), it does not guarantee the per-BS power budget constraint (i.e., the total power use over all channels should not be greater than P_{max}^M or P_{max}^S) is also satisfied. For example, for $n, n+1 \in \mathcal{N}^M$, having $P_{0,n}^m = P_{max}^M$ and $P_{0,n+1}^m = P_{max}^M$ at the same time does not violate the power budget constraint of each channel manager. However, that is not a feasible solution because the sum transmission power over channels cannot exceed P_{max}^M. Therefore, the upper level entity should check whether the sum power constraint is violated or not. To this end, we use a simple policy for the upper level entity that if the aggregate power budget constraint is violated, let all active channel managers use the same transmission power.[5] To be specific, for each MBS m of which sum power constraint is violated, let each channel manager with an associated UE allocate the transmission power in the following manner, $P_{0,n}^m = P_{max}^M/|\mathcal{U}_0^m|$, where $|\cdot|$ is the cardinality of a set $\cdot$. For each active SBS s of which sum power constraint is violated, let each channel manager with an associated UE use the power as $P_{s,n}^m = P_{max}^S/|\mathcal{U}_s^m|$. On the other hand, as long as the sum power constraint is satisfied, the upper level entity does not interrupt the power update procedures at lower-level channel managers.

E. Convergence

By using the analytical framework for convergence presented in [39], we prove that the proposed distributed power control algorithms in Eq. (13) and Eq. (16) converge to their corresponding optimum. Since both algorithms share the same structure and do not interfere with each other, we prove the convergence for an MBS m, i.e., Eq. (13). However, the following proof can be easily applied to the case for any SBS $s \in \mathcal{S}^m$, i.e., Eq. (16). To begin with, if the feasible power region is empty, i.e., if the power allocation problem P. 6 is infeasible, the transmission power for each channel converges (abruptly) to $P_{max}^M/|\mathcal{U}_0^m|$. This is because each channel manager is forced to use the equal transmission power when the sum power constraint is violated. In order to show convergence of the proposed scheme for the case of having a non-empty feasible power region, what follows is to transform the proposed power control method to the form of *interference function*, which is one of the key components in convergence analysis proposed in [39].

Since we consider only the case that the feasible power region is nonempty, we can simplify P. 6 by ignoring both constraints (6d) and (6e). After rearranging (6c), then, we get the following problem P. 17.

$$\min_{\mathbf{P}_0^m \succeq 0} \sum_{n\in\mathcal{N}^M} P_{0,n}^m \tag{17a}$$
$$\text{s.t.} \quad \eta_{u,n}^m \geq \eta_{thr} \cdot X_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m \tag{17b}$$
$$\eta_{u,n}^m \geq 2^{\tilde{r}_u^m \cdot X_{u,n}^m} - 1, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m. \tag{17c}$$

Since the optimality of P. 17 is achieved when among the two constraints, (17b) and (17c), the one that requires a higher transmission power is satisfied with equality, we can rewrite the problem as follows.

$$\min_{\mathbf{P}_0^m \succeq 0} \quad 0 \tag{18a}$$
$$\text{s.t.} \quad \eta_{u,n}^m = q_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m, \tag{18b}$$

where $q_{u,n}^m = \max\{\eta_{thr} \cdot X_{u,n}^m, 2^{\tilde{r}_u^m \cdot X_{u,n}^m} - 1\}$. The problem is always feasible by assumption, and the optimal power for each active channel is given by solving the equality constraint

[5] Assigning the same power to all active channels provides close-to-optimal performance [38].

Eq. (18b). Therefore, we can find the power update method directly from P. 18 after plugging in the SINR expression, Eq. (5a), to $\eta^m_{u,n}$. Then, the distributed power update method for an active channel n of MBS m which is associated with UE u becomes: $P^m_{0,n}[t+1] = \frac{q^m_{u,n}}{G^{0,m}_{u,n}}(\sum_{m' \neq m \in \mathcal{M}} G^{0,m'}_{u,n} \cdot P^{m'}_{0,n}[t] + \Delta f \cdot N_0)$. It is worth mentioning that this is equivalent to the power update method in Eq. (13) provided the feasible power region is nonempty and both the channel gain and the amount of interference are reported from UE.

According to [39], an iterative power update method, in general, is given by $\mathbf{p}[t+1] = \mathbf{I}(\mathbf{p}[t])$, where $\mathbf{I}(\cdot)$ is *interference function*.[6] We use $I^m_{0,n}(\cdot)$ to indicate the interference function for an active channel n of MBS m. The interference function is *standard* if it satisfies *positivity*, *monotonicity* and *scalability* properties for all nonnegative power vectors. Also, we use an overloaded notation $\mathbf{P}_0$ to indicate the transmission power of all MBSs. The positivity property, $I^m_{0,n}(\mathbf{P}_0) > 0$, is always satisfied because of the strictly positive background noise—even when $\mathbf{P}_0 = \mathbf{0}$, we have $\frac{q^m_{u,n}}{G^{0,m}_{u,n}} \Delta f \cdot N_0 > 0$. The interference function also satisfies the monotonicity property, i.e., if $\mathbf{P}^+_0 \succeq \mathbf{P}_0$, then $I^m_{0,n}(\mathbf{P}^+_0) \geq I^m_{0,n}(\mathbf{P}_0)$. Let $\mathbf{P}^+_0 = (1+\epsilon)\mathbf{P}_0$ for $\epsilon \geq 0$. Then, we have

$$\begin{aligned} I^m_{0,n}(\mathbf{P}^+_0) &= I^m_{0,n}((1+\epsilon)\mathbf{P}_0) \\ &= I^m_{0,n}(\mathbf{P}_0) + \frac{q^m_{u,n}}{G^{0,m}_{u,n}}(\epsilon \sum_{m' \neq m \in \mathcal{M}} G^{0,m'}_{u,n} \cdot P^{m'}_{0,n}) \\ &\geq I^m_{0,n}(\mathbf{P}_0), \end{aligned}$$

from which we can conclude that the monotonicity property is always satisfied. Finally, the positivity property and convexity of the interference function imply scalability, i.e., for all $\alpha > 1$, $\alpha I^m_{0,n}(\mathbf{P}_0) > I^m_{0,n}(\alpha \mathbf{P}_0)$.

Since the interference function $I^m_{0,n}(\cdot)$ satisfies the three properties, the proposed power update method is called *standard power control algorithm* [39]. Due to the convexity of the problem P. 17 (or P. 6), there exists an optimal power allocation vector, meaning that the proposed power update method has a fixed point. Then, the fixed point is unique by [39, Theorem 1]. Finally, by using [39, Theorem 2] we conclude that the proposed power update method converges to a unique fixed point for any initial power vector as long as the feasible power region is not empty. ■

F. Overall Procedure

In this section, the overall procedures of the proposed scheme is given, i.e., the BS association and channel assignment in Stage 1 and the power allocation in Stage 2, as a summary of the current section. At the beginning of Stage 1, all BSs transmit pilot signals over the entire channels to which they have an access. UE u senses the signal, calculates the per-channel cost, and transmits the cost vector $\mathbf{c}^m_u$ to MBS m. Then, MBS m determines the UE-BS association and channel assignment by running Algo. 1. The decision made by MBS m is broadcasted to all SBSs ($\forall s \in \mathcal{S}^m$) and all UEs ($\forall u \in \mathcal{U}^m$). Each active MBS channel manager with an associated UE runs Eq. (13) to determine the downlink power, and the UE sends the measured channel gain and interference back to the channel manager. This power allocation procedure iterates until the change of the power becomes less than the given threshold.Each active SBS channel manager with an associate UE runs the same procedure except that it runs Eq. (16) to determine the downlink transmission power. While each active channel manager tries to determine the transmission power, each BS checks if the sum power exceeds the power budget. If it does, the BS stops all active channel managers and lets them use the same power, $P^m_{0,n} = P^M_{max}/|\mathcal{U}^m_0|$ (in case of MBS) and $P^m_{s,n} = P^S_{max}/|\mathcal{U}^m_s|$ (in case of SBS) for downlink communication. If it does not, the BS waits until all active channel managers finish their power allocation procedures.

IV. Evaluation

We implemented and simulated the proposed algorithm along with others for comparison on top of MATLAB [40] and CVX [41]. The following Section IV-A describes the network configurations and parameter settings which are common to all scenarios considered in this section. In Section IV-B we show that for Stage 1, the optimality gap between the proposed solution (i.e., Algo. 1) and the optimal BS association and channel assignment (i.e., P. 2) is small. What follows is the performance evaluation and comparison of the proposed scheme to others in terms of the power consumption on different networks, i.e., single-cell, small-scale and large-scale networks, in Section IV-C, Section IV-D and Section IV-E, respectively.

For comparison to the optimal solution, we introduce a new metric, $D_X(\cdot)$, to measure the difference in the Stage 1 decision between a certain method and the optimal solution, which is defined as follows: $D_X(\texttt{<method>}) = ||X^*_{optimal} - X^*_{method}||_1$, where $||Y||_1 = \sum_{\forall y \in Y} |y|$. Here, $X^*_{optimal}$ is the optimal solution found by solving P. 2, whereas X^*_{method} is the optimal solution found by solving the corresponding problem for `<method>`. In other word, $D_X(\texttt{<method>})$ counts the number of entries that do not match between the two solutions.

In addition, we have implemented one more scheme, called SSSF (Strongest Signal Strength First) for comparison. In contrast to the proposed method which considers both the channel gain and the service demand, SSSF takes only the signal strength into account when making BS association and channel assignment. It is easy to implement SSSF or any similar variations due to the general structure of the Stage 1 problem. In contrast to the proposed method which minimizes the cost values, SSSF maximizes the sum of the benefit which is equal to the channel gains. After replacing $\mathbf{c}^m$ with the benefit, we can simply replace the objective function P. 2 with the benefit-sum maximization problem. We have directly solved SSSF by using the MATLAB (M)ILP solver which uses Branch-and-Bound algorithm.

A. Network Configuration

There are M MBSs that are regularly deployed with keeping the inter-cell distance of 600 m among adjacent ones. Each

[6] Note that the notation $\mathbf{I}$ in Section III-E is different from the one in Section III-C.

MBS has 300 m of coverage, and is overlaid by S indoor SBSs and U UEs. We have used a Uniform distribution for locating SBSs and UEs. A UE is located indoor if it is placed within the coverage of a SBS which is 30 m. The QoS requirement of each UE is randomly drawn from a Uniform distribution. The total BS transmission power of MBS and SBS is 46 dBm (40 W) and 20 dBm (100 mW), respectively [33].

The channel model from [33] is used, which includes the distance dependent path-loss, penetration loss (when applicable), multipath fading and lognormal shadowing. The path-loss between a BS and a UE is listed below. The unit of path-loss is dB, and R is the distance between two entities in the unit of meter.

- MBS and an indoor UE:
  $15.3 + 37.6\log(R) + L_{ow}$,
- MBS and an outdoor UE:
  $15.3 + 37.6\log(R)$,
- SBS and its associated UE:
  $38.46 + 20\log(R)$, and
- SBS and an outdoor UE:
  $\max\{38.46 + 20\log(R), 15.3 + 37.6\log(R)\} + L_{ow}$,

where L_{ow} is the penetration loss of an outdoor wall, which is 20 dB. In case of the path-loss between an indoor UE and an SBS located in a different building, the penetration loss gets doubled. The Rayleigh fading model is used to capture the multipath effect, and the standard deviation of lognormal shadowing is as follows.

- MBS and an indoor UE: 10 dB,
- MBS and an outdoor UE: 10 dB,
- SBS and its associated UE: 4 dB, and
- SBS and an outdoor UE: 8 dB.

In addition, for a fair comparison to [27] we set $\Delta f = 180$ kHz and per-Hz noise power $N_0 = 10^{-13}$ W in accordance with the parameters declared therein.

B. Optimality Gap in Stage 1

As aforementioned in Section III-C, the Algo. 1 iteratively solves P. 3 and recovers binary solutions instead of directly solving P. 2 to lower the computational complexity. Due to the relaxation on binary variables, Algo. 1 may yield a suboptimal solution to P. 2 which possibly affects the power consumption in the subsequent Stage 2 for power allocation. In order to check by how much the solution of Algo. 1 is deviated from the optimal solution to P. 2, we have implemented and solved P. 2 by using the MATLAB (M)ILP solver which uses Branch-and-Bound algorithm. Each data point in both Fig. 1 and Fig. 2 is an average of 20 runs of randomly-generated scenarios, where there are 4 SBSs on an MBS. The number of UEs in a macrocell is set to 10, 20, $\cdots$, 50. Also, 95% of the confidence interval is marked on each data point in Fig. 1.

Fig. 1 shows the (normalized) minimum cost found by running Algo. 1 and P. 2, referred to as Proposed and Optimal, respectively, in the figure. For each different number of UEs on a macrocell, the proposed method results in a close-to-optimal objective value. In order to take a closer look at the difference in the two objective values, Fig. 2 shows the error ratio $e = |\hat{p}^* - p^*|/p^*$, where $\hat{p}^*$ and p^* are the normalized minimum cost found by running Algo. 1 and solving P. 2, respectively. As it can be seen in Fig. 2 the error ratio becomes stable as the number of UEs increases and does not exceed 0.008. That is, the proposed Algo. 1 yields a sub-optimal solution to P. 2 with a small optimality gap.

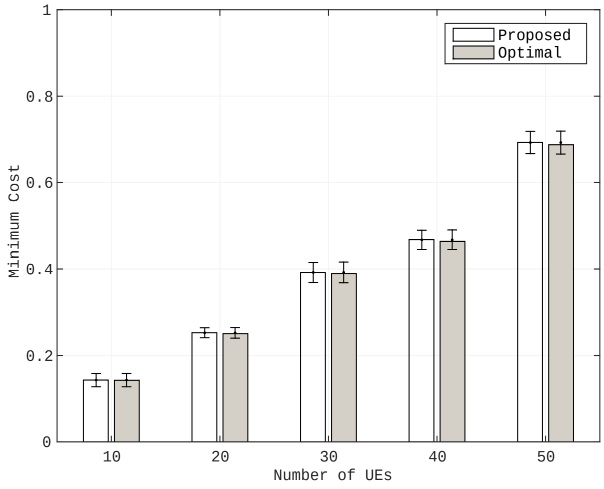

Figure 1. Minimum cost in Stage 1 for the optimal and the proposed BS/channel assignment method.

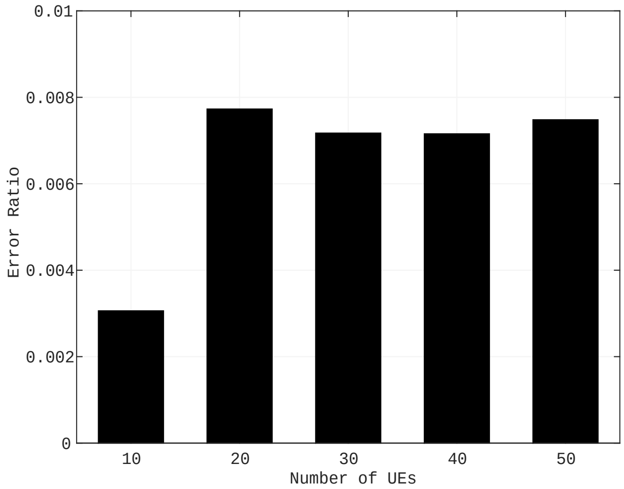

Figure 2. Error ratio of the minimum cost in Stage 1 for the proposed BS/channel assignment method.

In what follows, we show the power consumption of the proposed method along with others for comparison, and show the effect of the sub-optimality on the power consumption.

C. Single-Cell Networks

In addition to comparing to the optimal solution and SSSF, we compare the performance of the proposed method to [27], which is denoted by Abdelnasser in Fig. 4 and Fig. 5. In contrast to the proposed scheme which assumes an independent channel deployment between different types of BSs, [27] shares all available channels between an MBS and all SBSs

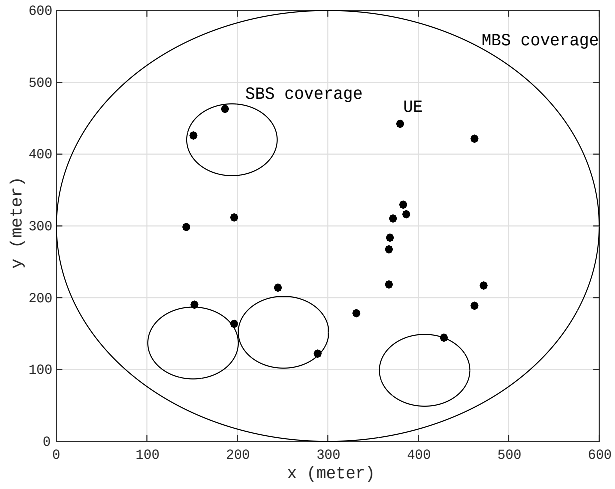

Figure 3. Network scenario for a single-cell network.

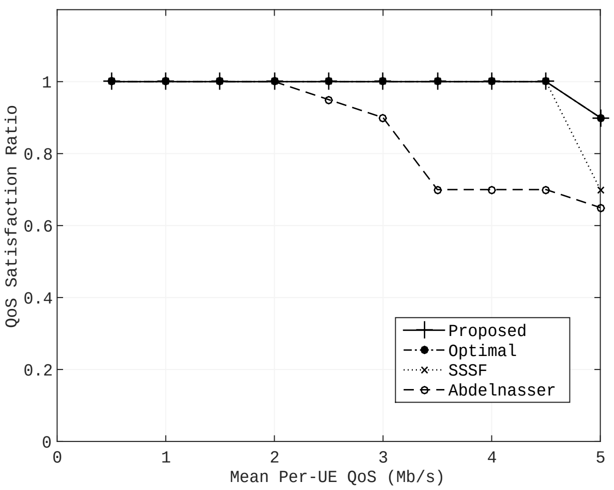

Figure 5. QoS satisfaction ratio for a single-cell network.

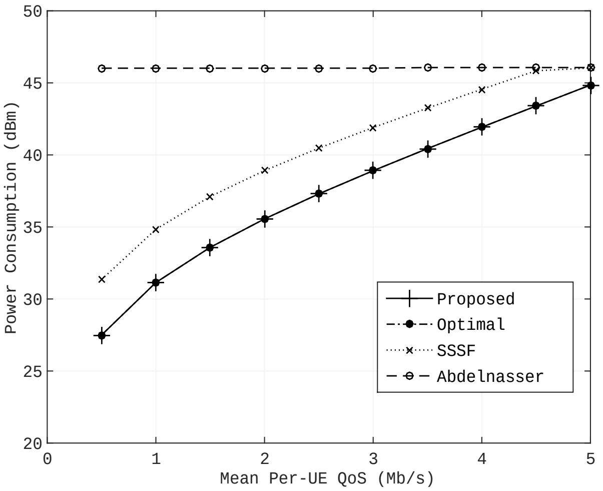

Figure 4. Overall power consumption for a single-cell network.

therein, called co-channel deployment. The proposed resource scheduling scheme in this paper allocates an optimal power to each channel by taking both the channel gain and the QoS requirement of an associated UE into account, while [27] allocates an equal power to all channels in use. The proposed association scheme in this paper allows UEs to be dynamically offloaded to SBSs for capacity enhancement (i.e., open access mode), whereas [27] does not (i.e., closed access mode). Since the work in [27] considers only a single MBS, we set up a similar network as Fig. 3, where there is a single MBS on the network with 4 SBSs and 20 UEs therein. Please note that due to this limitation, [27] is not used for performance comparison in the following Section IV-D and Section IV-E. In this simulation, we have $D_X(\texttt{proposed}) = 0$ and $D_X(\texttt{SSSF}) = 6$.

1) Power Consumption: Fig. 4 shows the overall power consumption, i.e., the total power used by an MBS and SBSs on the network, for the four different schemes. As it can be seen in Fig. 4, the work in [27] uses more power than the proposed method as well as Optimal and SSSF. In contrast to the proposed method in this paper that dynamically manages the interference, [27] takes a *conservative* approach. In [27], when a UE is associated with an MBS, the MBS calculates the maximum allowable interference on the allocated channel for the UE, and then assigns the maximum power to the channel which is P^M_{max} divided by the number of channels in use. On the other hand, the proposed scheme in this paper as well as both Optimal and SSSF allocates the minimum power to each channel while satisfying both SINR and QoS requirements for the associated UE. Therefore, our proposed work much outperforms [27] in terms of the power consumption especially when the amount of downlink traffic is small. Since we have $D_X(\texttt{proposed}) = 0$, the proposed scheme results in the optimal solution. SSSF is also able to dynamically adjust the transmission power, and thus, its power usage gradually increases as the average service demand increases. However, our proposed scheme consumes less power than SSSF. That is, considering both channel gain and QoS requirements results in a more power-efficient solution than taking only the signal strength into account.

2) QoS Satisfaction Ratio: The Fig. 5 shows the QoS satisfaction ratio which is the ratio of the number of UEs with their QoS satisfied to the total number of active UEs on the network. As it can be seen in the figure, the satisfaction ratio of the work in [27] starts to drop when the mean QoS becomes larger than 2 Mb/s. On the other hand, the other methods successfully satisfy the QoS requirements of all UEs until the mean QoS is 4.5 Mb/s. Due to the lack of freedom in controlling the downlink transmission power, the work in [27] always allocates the fixed transmission power to all active channels. This inflexibility in power allocation may not be efficient, because it allocates more than necessary amount of power to the UEs with high channel gains, while failing to satisfy the QoS requirement of UEs with low gains. The proposed method and the Optimal, on the other hand, has a higher level of freedom in power control than [27], because

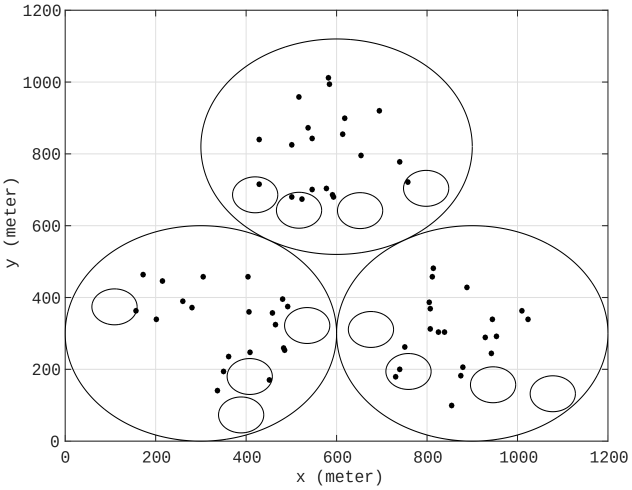

Figure 6. Network scenario for a small-scale network.

each channel manager allocates the minimum power level that satisfies both SINR and QoS requirements of the associated UE. When the average per-UE QoS is 5 Mb/s, the QoS satisfaction ratio of SSSF drops sharply, while it is not the case for both the proposed and optimal scheme even though both experience a small amount of degradation.

D. Small-Scale Networks

We have evaluated the proposed method on a small-sized network where there are 3 MBSs on the network as shown in Fig. 6. The locations of these three MBSs form an equilateral triangle, meaning that the distance from any MBS to either of the rest two is the same. Each MBS is overlaid by 4 SBSs and 20 UEs. In this simulation, we have $D_X(\texttt{proposed}) = 0$ and $D_X(\texttt{SSSF}) = 18$. Therefore, the performance of the proposed scheme will be exactly same as that of Optimal.

1) Power Consumption: The Fig. 7 shows the power consumption of the three methods, the proposed, optimal and SSSF, with respect to different mean per-UE QoS. The overall power use is the sum of transmission power used by all macro and smallcell BSs on the network. Although it is not shown in the figure, the difference between the overall power use and the aggregate MBS power use is trivial, meaning that MBSs use most of the power consumed in the network. This is because an MBS associates with much larger number of UEs than SBSs due to the long transmission range and a larger power budget. Thus, a UE associated with an MBS may have a small channel gain, making the MBS use a high transmission power to satisfy the UE's SINR and QoS requirement. On the other hand, an SBS has a small number of associated UEs with a short distance to each. Therefore, it does not need much power to satisfy the associated UE's QoS demand and SINR requirement.

The overall power consumption becomes saturated when the mean per-UE QoS demand is approximately 5 and 6 Mb/s, respectively, for SSSF and both the proposed and optimal schemes.

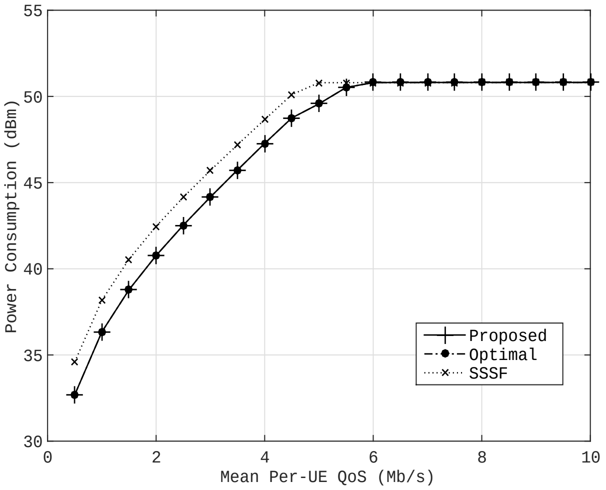

Figure 7. Overall power consumption for a small-scale network.

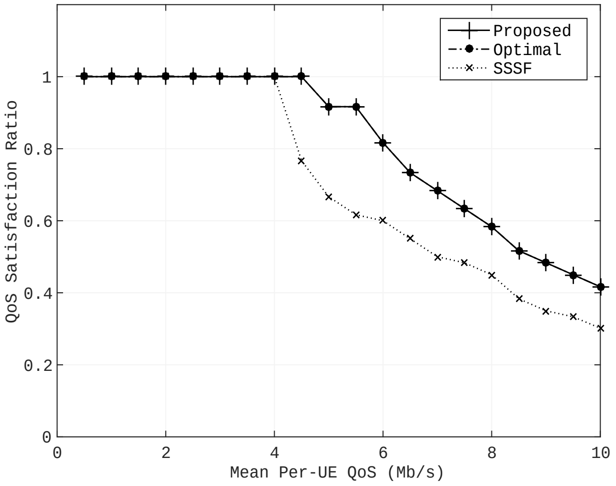

Figure 8. QoS satisfaction ratio for a small-scale network.

2) QoS Satisfaction Ratio: QoS satisfaction ratio is the number of UEs with their QoS satisfied to the total number of active UEs on the network. As it can be seen in Fig. 8, the QoS requirements of all UEs are fully satisfied when the mean per-UE QoS is equal to or less than 4 or 4.5 Mb/s, respectively, for SSSF or both the proposed and optimal. Then, the QoS satisfaction ratio drops as the per-UE demand becomes larger. It is noteworthy that for the first one or two drops of the ratio, SSSF shows a steeper decline than the rest two. Since it already uses much power when the mean per-UE QoS is 4 Mb/s, further increase in QoS causes a significant drops in the QoS satisfaction ratio. The increase in the QoS requirement will eventually let all BSs use the maximum transmission power, which increases the interference level. Failing in achieving the satisfaction ratio of 1 means there is at least one BS whose total power budget constraint is violated. Note that the violation of the power budget constraint makes a BS use an equal power allocation for all active channels. Since

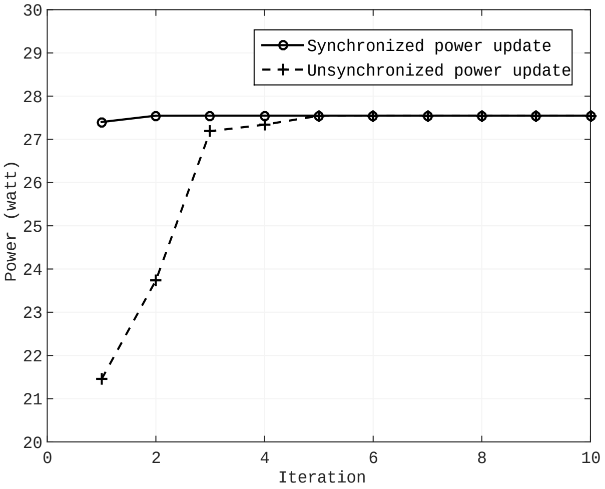

Figure 9. Convergence of the iterative power allocation procedure for a small-scale network.

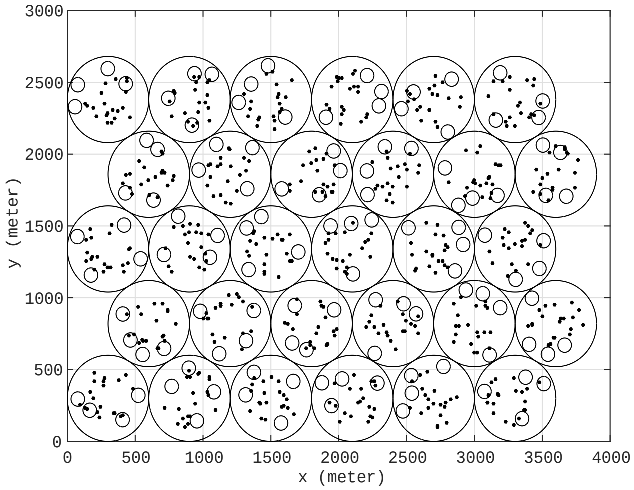

Figure 10. Network scenario for a large-scale network.

the equal power assignment takes away the freedom in power control from a BS, any further increase of QoS requirements will yield more UEs with their QoS unsatisfied. UEs in a SSSF network suffer much more QoS degradation as the average QoS demand increases compared to both the proposed and the optimal schemes.

3) Convergence: The speed of convergence determines whether the proposed algorithm is suitable for an online processing or not. We have evaluated the speed of convergence with two different setting as follows, while keeping the rest network configurations the same. One is *synchronized* power update and the other is *unsynchronized*. In the synchronized power update setting, all active channel managers update their power at the same time. In other words, on iteration t it is guaranteed that all the active channel managers on the network have fished their power update procedures for the previous $t-1^{st}$ iteration. On the other hand, the unsynchronized power update does not assume the synchronized power update. For example, when a channel manager runs its t^{th} power update, it is possible that there are some channel managers that have not finished their $t-1$ iterations (or even the ones before the $t-1^{st}$ iteration). In order to emulate the unsynchronized updates, we have added some random delay into the power update method. The Fig. 9 shows the convergence behaviors of the power update method under different settings. As expected, the unsynchronized setting requires more iterations. In this small-scale network, every MBS has two effective interfering cells,[7] both of which are 600 m away. Therefore, when the update procedures are synchronized, all BSs will converge at the same time on a small-scale network. Note that this is not the case for a large-scale network where each cell is exposed to the different number of effective interfering cells.

E. Large-Scale Networks

We also carried out an evaluation on a large-scale network where there are 30 MBSs, each of which is overlaid by 4 SBSs and 20 UEs. MBSs are placed in a shape of a bee hive where there are 5 rows of MBSs with 6 MBSs per row as in Fig. 10. The rest configurations remain the same as before. In this simulation, we have $D_X(\texttt{proposed}) = 22$ and $D_X(\texttt{SSSF}) = 358$. The proposed method has failed to find the optimal solution in BS association and channel assignment for this network for having a nonzero value of $D_X(\texttt{proposed})$. However, since the difference is not significant, the power consumption in Stage 2 is expected not to be deviated much from that of the optimal method.

1) Power Consumption: The power consumption of BSs on a large-scale network is illustrated in Fig. 11, which has a similar trend to Fig. 7. It is worth mentioning that although the proposed method has a (trivially) different Stage 1 solution than that of the optimal method, it is hardly seen on the power consumption. In both small- and large-scale cases, the overall power use of both the proposed and optimal becomes saturated around 6 Mb/s of the mean QoS. Also, both methods outperform SSSF in terms of power consumption, indicting that considering both channel gain and QoS requirement is more energy-efficient than the one considering only the received signal strength.

2) QoS Satisfaction Ratio: Each cell on the large-scale network experiences more interference than that on the small-sized network for the increased number of effective interfering cells. Fig. 12 shows the QoS satisfaction ratio of UEs on the large-sized network. As it can be seen in the figure, the ratio starts to drop when the mean QoS becomes larger than 2 and 3.5 Mb/s, respectively, for SSSF and both the proposed and optimal, which was not the case on the small-scale network. This is mainly because of the increased level of interference on the large-scale network. Having more effective interfering cells on the network increases the level of interference to each cell, and thus results in a lower energy efficiency. The performance

[7] A cell is an effective interfering cell to another if the interfering cell is close enough to the interfered cell. For example, if an interfering cell is 600 m away, it is an effective interfering cell. However, a cell which is 600 km away is not an effective interfering cell, because the interference from the cell is too weak.

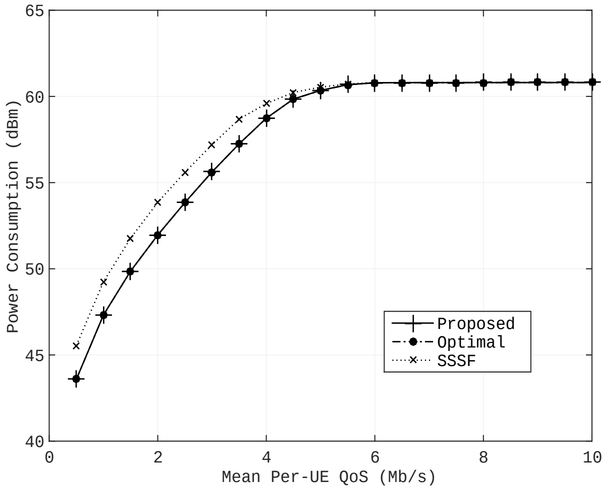

Figure 11. Overall power consumption for a large-scale network.

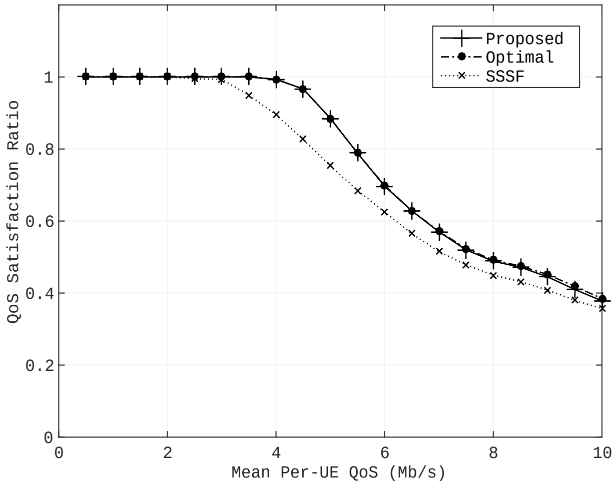

Figure 12. QoS satisfaction ratio for a large-scale network.

degradation of the proposed method for having a non-optimal solution in Stage 1 can be seen in Fig. 12. As the average QoS requirement becomes larger than 8 Mb/s, the proposed method results in a slightly lower QoS satisfaction ratio than the optimal method, but the degradation is trivial.

3) Convergence: Due to the increased number of interfering cells, the iterative power update procedure on the large-scale network takes a couple of more steps to converge as shown in Fig. 13. Still, the network reaches convergence in 6 or 7 iterations even when the power update is not synchronized. This fast convergence behavior on the large-scale network proves that the proposed scheme scales well with the network size, making it suitable for an online resource scheduling.

V. Conclusion

In this paper, we have proposed a distributed and energy-efficient association and resource scheduling scheme for two-tier HetNets. We have formulated an optimal user association problem and then proposed a resource allocation algorithm in such a way that they can be solved efficiently by an iterative, distributed method. To be specific, we have formulated an optimal user association and channel assignment problem for Stage 1 and then applied a relaxation and an iterative adjustment method so as to make the problem tractable and low-complex. In addition, we have transformed the proposed power assignment problem into a set of lightweight distributed procedures by using the decomposition structure for Stage 2. The comparison results and the evaluation studies on small-/large-scale networks show that the proposed scheme maintains a low power consumption while satisfying users' QoS requirements with a low computational load, which proves that the proposed scheme can be used for an online resource scheduling for HetNets.

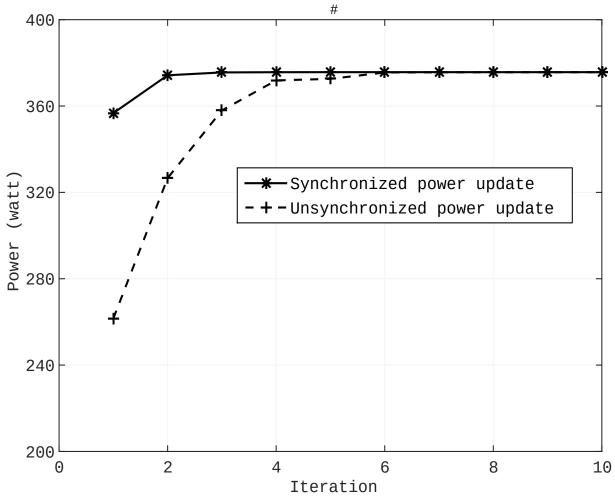

Figure 13. Convergence of the iterative power allocation procedure for a large-scale network.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019," Cisco Systems, Inc., San Jose, CA, Cisco White Paper, Feb. 2015.

[2] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, "Energy efficiency in the future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 2, pp. 223–244, July 2011.

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

[4] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.

[5] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the horizon: Key challenges for the radio-access network," *IEEE Vehecular Technology Magazine*, vol. 8, no. 3, pp. 47–53, July 2013.

[6] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 106–112, Apr. 2014.

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

[8] T. Q. S. Quek, G. de la Roche, I. Guvenc, and M. Kountouris, *Small cell networks: deployment, PHY techniques, and resource management*, Cambridge, United Kingdom: Cambridge University Press, 2013.

[9] 3rd Generation Partnership Project (3GPP), Evolved universal terrestrial radio access (E-UTRA) and evolved universal universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2, TS 36.300, Release 13, June 2016.

[10] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond [Accepted From Open Call]," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 154–160, July 2013.

[11] I. Ashraf, F. Boccardi, and L. Ho, "SLEEP mode techniques for small cell deployments," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, Aug. 2011.

[12] Y. Li, M. Sheng, C. W. Tan, Y. Zhang, Y. Sun, X. Wang, Y. Shi, and J. Li, "Energy-efficient subcarrier assignment and power allocation in OFDMA systems with max-min fairness guarantees," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3183–3195, Sep. 2015.

[13] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," vol. 14, no. 1, pp. 1–14, Jan. 2015.

[14] Y.-P. Zhang, S. Feng, and P. Zhang, "Adaptive cell association and interference management in LTE-A small-cell networks," in *IEEE Vehicular Technology Conference (VTC Fall)*, Las Vegas, NV, 2013, pp. 1–6.

[15] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. of IEEE International Conference on Computer and Communications*, Turin, 2013, pp. 998–1006.

[16] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, Dec. 2014.

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

[18] W. Wang, X. Wu, L. Xie, and S. Lu, "Femto-matching: efficient traffic offloading in heterogeneous cellular networks," in *Proc. of IEEE International Conference on Computer and Communications*, Kowloon, 2015, pp. 325–333.

[19] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1100–1113, June 2014.

[20] V. N. Ha and L. B. Le, "Fair resource allocation for OFDMA femtocell networks with macrocell protection," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1388–1401, Oct. 2014.

[21] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for OFDMA femtocell networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 342–355, Dec. 2014.

[22] W. C. Cheung, T. Q. S. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE Journal on Selected Area in Communications*, vol. 30, no. 3, pp. 561–574, Apr. 2012.

[23] W. Bao and B. Liang, "Radio resource allocation in heterogeneous wireless networks: A spatial-temporal perspective," in *Proc. of IEEE International Conference on Computer and Communications*, Kowloon, 2015, pp. 334–342.

[24] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, Dec. 2013.

[25] V. Chandrasekhar and J. G. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Transactions on Communications*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[26] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2027–2038, May 2015.

[27] A. Abdelnasser, E. Hossain, and D. I. Kim, "Tier-aware resource allocation in OFDMA macrocell-small cell networks," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 695–710, Feb. 2015.

[28] Y. Li, T. Jiang, M. Sheng, and Y. Zhu, "QoS-aware admission control and resource allocation in underlay device-to-device spectrum-sharing networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 11, pp. 2874–2886, Nov. 2016.

[29] K. Son, S. Lee, Y. Yi, and Song Chong, "REFIM: A practical interference management in heterogeneous wireless access networks," *IEEE Journal on Selected Areas in Communications,* vol. 29, no. 6, pp. 1260–1272, Jun. 2011.

[30] Y. Li, M. Sheng, Y. Sun, and Y. Shi, "Joint Optimization of BS Operation, User Association, Subcarrier Assignment, and Power Allocation for Energy-Efficient HetNets," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3339–3353, Dec. 2016.

[31] G. Bacci, E. V. Belmega, P. Mertikopoulos, and L. Sanguinetti, "Energy-aware competitive power allocation for heterogeneous networks under QoS constraints," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4728–4742, Apr. 2015.

[32] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admission game for uplink user association in wireless small cell networks," in *Proc. of IEEE International Conference on Computer and Communications*, Toronto, ON, 2014, pp. 1096–1104.

[33] 3rd Generation Partnership Project (3GPP), Evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects, TR 36.814, Release 9, Mar. 2010.

[34] J.-C. Lin, T.-H. Lee, and Y.-T. Su, "Power control algorithm for cellular radio systems," *IET Electronics Letters*, vol. 30, no. 3, pp. 195–197, Feb. 1994.

[35] J. Zander, "Distributed cochannel interference control in cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 3, pp. 305–311, Aug. 1992.

[36] M. Andersin, Z. Rosberg, and Z. Zander, "Gradual removals in cellular pcs with constrained power control and noise," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Toronto, ON, 1995, vol. 1, pp. 56–60.

[37] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[38] D. Tse and P. Viswanath, *Fundamentals of wireless communication*, Cambridge, United Kingdom: Cambridge University Press, 2005.

[39] R. D. Yates, "A framework for uplink power control in cellular radio systems", *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.

[40] Matlab, MathWorks. Inc., Natick, MA, http://www.mathworks.com.

[41] CVX: Matlab software for disciplined convex programming, version 2.0, CVX Research, Inc., http://cvxr.com/cvx.

Taewoon Kim received the B.S. degree in Computer Science and Engineering from Pusan National University, Republic of Korea, in 2008 and the M.S. degree in Information and Mechatronics from Gwangju Institute of Science and Technology, Republic of Korea, in 2010. He is currently pursuing the Ph.D. degree in computer engineering at Iowa State University, Ames, IA. From 2010 to 2013, he was a Research Engineer at the Telecommunications Technology Association, Republic of Korea. His research interest includes network modeling, optimization and protocol design on wireless networking systems, such as WLAN, sensor networks, cellular networks and heterogeneous networks.

J. Morris Chang received the MS and PhD degrees in computer engineering from North Carolina State University, Raleigh, NC. He joined University of South Florida, Tampa, FL, in August, 2016. He was on the faculty of the Department of Electrical Engineering at Rochester Institute of Technology, Rochester, NY, from 1993 to 1995, the Department of Computer Science at the Illinois Institute of Technology, Chicago, IL, from 1995 to 2001, and the Department of Electrical and Computer Engineering at Iowa State University, Ames, IA, from 2001 to 2016. He received the IIT University Excellence in Teaching Award in 1999. His research interests include cyber security, wireless networks, energy-aware computing, and object-oriented systems. His research projects have been supported by US National Science Foundation (NSF), US Defense Advanced Research Projects Agency (DARPA), and Altera. Currently, he is a handling editor of the Journal of Microprocessors and Microsystems, and the Associate Editor-in-Chief of IEEE IT Professional. He is a senior member of the IEEE.