Joint Frame Design, Resource Allocation and User Association for Massive MIMO Heterogeneous Networks With Wireless Backhaul

Mingjie Feng, *Student Member, IEEE*, Shiwen Mao, *Senior Member, IEEE*,
and Tao Jiang, *Senior Member, IEEE*

***Abstract*— In this paper, we investigate the problem of frame design, resource allocation, and user association in a massive multiple input multiple output (MIMO) heterogeneous network (HetNet) with wireless backhaul (WB) and linear processing. The objective is to maximize the sum downlink rate of all users, subject to constraints on data rates of WBs and fairness-aware constraints. Such a problem is formulated as an integer programming problem with both coupled variables and coupled constraints. We first develop a centralized scheme in which we decompose the original problem into two subproblems and iteratively solve them until convergence to achieve a near-optimal solution. We then propose a distributed scheme by formulating a repeated game among all users and prove that the game converges to a Nash Equilibrium. Simulation studies show that the proposed schemes are adaptive to different network scenarios and traffic patterns, and achieve considerable gains over several benchmark schemes.**

***Index Terms*— 5G Wireless, massive MIMO, HetNet, cross-layer optimization, wireless backhaul.**

I. Introduction

WITH the fast growing popularity of smart mobile devices and the explosion of data-intensive services, the wireless system is expected to provide a 1000× mobile data rate in the near future. To support such high data rate with limited spectrum, aggressive spectrum reuse must be realized to achieve high spectral efficiency. To this end, *massive MIMO* (Multiple Input Multiple Output) and *small cell* are recognized as two key technologies for emerging 5G wireless systems [2]. Massive MIMO refers to a cellular system with more than 100 antennas equipped at the base station (BS), which serves multiple users with the same time-frequency resource [3]. A massive MIMO system can dramatically improve the energy and spectral efficiency compared to traditional wireless systems due to highly efficient spatial multiplexing [4]–[6]. Small cell deployment, which forms a heterogeneous network (HetNet), is another efficient approach to enhance spectral efficiency. Due to the short transmission range, high signal to noise ratio (SNR) and dense spectrum reuse can be achieved, resulting in significantly improved network capacity.

As an integration of these two techniques, *massive MIMO HetNet* has drawn considerable attention recently [7]–[11], where the macrocell BS (MBS) is equipped with a large number of antennas. The MBS and small cell BS's (SBS) collectively serve users in the cell. With such a network architecture, the MBS with massive MIMO can provide a good quality of service (QoS) to users located in the coverage holes of SBS's. Moreover, the SBS's can offload some traffic from the MBS so that the overhead and complexity of processing at MBS can be reduced, resulting in performance enhancement of users that are still served by the MBS.

With the expected massive deployment of small cells, connecting all SBS's to the core network directly with dedicated optical fiber may not be feasible due to significantly increased cost. Alternatively, the SBS's can be connected to the core network by transmitting data to the MBS through backhaul links. In this case, the design of backhaul system is an important issue of a HetNet. Although a massive MIMO HetNet can provide high data rate links between users and BS's, the transmissions between MBS and SBS's may become the bottleneck of the network. Without a reliable backhaul, the aggregated data rate of small cell user equipments (SUE) would be limited by the data rate of the backhaul link. For services with stringent delay requirements, the QoS of users may become unacceptable or even causing outages.

Most existing works have considered wired backhaul between SBS's and MBS, since a wired connection can support high data rate and it is more reliable in general. However, in a HetNet with large number of SBS's, wired connections to each SBS may not be cost-effective or even may be infeasible due to practical constraints. Moreover, the wired backhaul deployment may be highly inefficient when the wireless service provider needs to upgrade or extend the

Manuscript received June 24, 2017; revised October 31, 2017; accepted December 17, 2017. Date of publication December 29, 2017; date of current version March 8, 2018. This work was supported in part by the NSF under Grant CNS-1320664 and Grant CNS-1702957, in part by the Wireless Engineering Research and Engineering Center, Auburn University, in part by the National Science Foundation for Distinguished Young Scholars of China (NSFC) with Grant 61325004, and in part by the NSFC under Grant 61771216. This paper was presented at the IEEE SECON 2017, San Diego, CA, USA, June 2017. The associate editor coordinating the review of this paper and approving it for publication was M. Li. *(Corresponding author: Shiwen Mao.)*

M. Feng and S. Mao are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: mzf0022@auburn.edu; smao@ieee.org).

T. Jiang is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: tao.jiang@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2017.2787139

network. Thus, the *wireless backhaul* (WB) has the potential to play an increasingly important role in 5G networks due to its easy and fast deployment, flexibility, and low cost [12]–[14]. In fact, WB in a massive MIMO HetNet can be quite reliable with proper configurations, especially when massive MIMO are applied with *linear processing* techniques. From the perspective of an MBS, supporting a WB transmission is equivalent to serving a macrocell user equipment (MUE). With linear processing, e.g., maximum ratio combination (MRC) and maximum ratio transmission (MRT), the reception and precoding are based on linear functions of channel response matrices. When the number of antennas goes to infinity, the inner products of channel vectors of different links grow at a lower rate than that of the number of antennas, the interference between different WBs or MUEs can be averaged out [3]. Thus, the MBS can provide high data rate links to multiple WBs with simple linear processing techniques.

The use of WB in massive MIMO HetNet has drawn some attentions recently [15]–[18]. In [15], a joint user association and bandwidth allocation scheme was proposed to maximize the downlink sum logarithmic data rate in a massive MIMO HetNet with zero-forcing (ZF) at MBS. A comparison of three WB deployment strategies are presented in [16], namely complete time division duplex, zero division duplex, and zero division duplex with interference rejection. An analytical framework based on stochastic geometry was presented in [17] to study the WB performance in a massive MIMO HetNet with full-duplex small cells, and a closed-form expression of coverage probability was derived. In [18], the network architecture and feasibility issues of WB on the mm-wave band were investigated in a dense HetNet with massive MIMO.

Although these works presented several highly efficient approaches, optimal frame design on pilots, i.e., the number of symbols used for pilots in each frame, has not been considered. Here, a frame is defined as a time-frequency resource block and the size of each frame is determined by the coherence time and coherence frequency of all UEs. In each frame, a certain fraction of time is used to transmit symbols that are used as pilots, and these pilots are sent by MUEs and WBs to estimate their channel gains to the MBS. While existing works assume a fixed fraction of time dedicated for pilot, the pilot length, i.e., the number of symbols used for pilots in each frame, can be adaptive to the traffic pattern for performance enhancement. There is clearly a *trade-off* on pilot length here. As discussed, the WBs and MUEs are equivalent from the MBS's point of view. When the pilot length is large, more time is spent on channel estimation at MBS, and a large number of MUEs and WBs can be supported. Moreover, the MUEs and WBs can be allocated with more channels since there is enough time to estimate all these channels. However, as a large fraction of time is dedicated to pilots, the remaining time for data transmission is small, resulting in a low data rate. When the pilot length is small, the fraction of time for data is increased, but the MUEs and WBs are allocated with less number of channels, which limits the data rates of MUEs and WBs. With a small data rate for WBs, the aggregated data rates of SUEs are limited, resulting in a poor performance.

In this paper, we investigate the problem of joint frame design, resource allocation, and user association to maximize the downlink sum rate of all users under the WB and fairness constraints. We develop efficient centralized and distributed schemes to obtain the near-optimal solutions to the formulated problem. The main contributions of this paper are as follows.

- We consider joint pilot length optimization, resource allocation, and user association in a massive MIMO HetNet with WB and linear processing, and provide a rigorous problem formulation.
- We propose a centralized iterative algorithm. The original problem is decomposed into two subproblems and we iteratively solve them until convergence. The first problem is joint pilot length optimization and resource allocation for MUEs and WBs, and we employ a primal decomposition approach to obtain its optimal solution. The second problem is user association, and we obtain its near-optimal solution with a cutting plane approach. An iterative framework is designed to update the parameters of the two subproblems in each iteration to minimize the performance gap between the two problems and guarantee that all constraints are satisfied.
- We propose a distributed scheme by formulating a repeated game among all users, and prove that the game converges to a Nash Equilibrium (NE).
- The performances of the proposed schemes are compared with several benchmark schemes. The simulation results show that performance gains can be as much as more than 100% under certain circumstances.

In the remainder of this paper, we present the system model and problem formulation in Section II. The centralized and distributed schemes are presented in Sections III and IV, respectively. We discuss our simulation study in Section V. Section VII concludes this paper.

II. Problem Formulation

We consider a noncooperative multi-cell cellular system with focus on a tagged macrocell (denoted as macrocell 0). Macrocell 0 is a two-tier HetNet consisting of an MBS with massive MIMO (indexed by $j = 0$) and J single-antenna SBS's (indexed by $j = 1, 2, \ldots, J$). The payload data of SUEs is transmitted to the core network via WBs between the MBS and SBS's. Then, the reversed time division duplex (RTDD) scheme is a natural choice for the MBS and SBS's [15]. With RTDD, the uplink and downlink transmissions of MBS and SBS's are performed in a reversed pattern, so that an SBS can transmit uplink data to (receive downlink data from) the MBS, and transmit downlink data to (receive uplink data from) SUEs simultaneously. The RTDD scheme is easy to implement in a practical system since it does not require interference cancellation at SBS's. There are K single-antenna mobile users (indexed by $k = 1, 2, \ldots, K$). Each user can be served by either the MBS or an SBS. We define binary variables for user association as

$$x_{k,j} \doteq \begin{cases} 1, & \text{user } k \text{ is associated with BS } j \\ 0, & \text{otherwise,} \end{cases}$$
$$k = 1, 2, \ldots, K, \quad j = 0, 1, \ldots, J. \quad (1)$$

The spectrum band owned by the wireless service provider (WSP) is divided into N channels, and the bandwidth of each channel is defined to be the coherence bandwidth of massive MIMO terminals [21]. We assume the MBS adopts *linear processing* schemes with MRC at receiver and MRT at transmitter [3], [19]. From the point of view of MBS, a WB is equivalent to a user to be served. Thus, we can take advantage of the favorable properties of massive MIMO by serving all MUEs and WBs on a same set of channels. This way, they can be put into the beamforming groups on these channels. Due to the law of large numbers, the interference between any two links in a beamforming group can be averaged out. From the perspective of an SBS, a WB is also equivalent to a user to be served. However, since the SBS's are assumed to be equipped with single antenna, they cannot perform interference mitigation in the spatial domain or self-interference cancelation. Hence, orthogonal resources must be assigned between WBs and SUEs to avoid mutual interference. Consequently, we assume that a proportion of α of the whole bandwidth is allocated to WBs and MUEs, and the rest $(1-\alpha)$ is allocated to SUEs. Note that α needs to be consistent across all macrocells to avoid cross-tier interference, and it is predetermined by the service provider.

We assume both the bandwidth of each frame and the bandwidth of a channel equal to the coherence bandwidth of all MUEs and WBs, given as W_c. Then, each frame corresponds to a specific interval on a channel. The duration of a frame is T_c seconds, which equals to the coherence time of all MUEs and WBs. Thus, the channel gains are constant in a frame and each frame can be viewed as a *coherence block*. The interval of a symbol is T_s seconds, which consists of T_u seconds for useful symbols and $T_g = T_s - T_u$ seconds for guard interval. Let Δ_f be the spacing of subcarriers, then T_u is given as $T_u = 1/\Delta_f$. Within a coherence bandwidth, there are W_c/Δ_f subcarriers. Hence, the channel response is constant over $N_{\text{sm}} = W_c/\Delta_f$ consecutive subcarriers in each symbol. Let τ be the pilot length, i.e., the number of OFDM symbols dedicated for pilots in each frame. Then, the number of terminals that can be supported in each frame is τN_{sm}. Therefore, the total number of MUEs and WBs that can be served by the MBS on each channel within the interval of a frame is upper bounded by τN_{sm}.

Given the available spectrum band for MUEs and WBs, we define the following resource allocation indicators

$$a_{k,n} \doteq \begin{cases} 1, & \text{channel } n \text{ is allocated to MUE } k \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, 2, \ldots, K, \quad n = 1, \ldots, \alpha N. \tag{2}$$

$$b_{j,n} \doteq \begin{cases} 1, & \text{channel } n \text{ is allocated to SBS } j\text{'s WB} \\ 0, & \text{otherwise,} \end{cases} \quad j = 1, 2, \ldots, J, \quad n = 1, \ldots, \alpha N. \tag{3}$$

According to our analysis, we have

$$\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\text{sm}}, \quad n = 1, \ldots, \alpha N. \tag{4}$$

In a massive MIMO system, the effects of fast fading and noise vanish as the number of antennas goes to infinity; the only possible interference comes from the UEs that share the same pilot sequence [3]. Thus, the only factor that limits the performance of a massive MIMO system with linear processing is pilot contamination. For user k connecting to the MBS in macrocell 0, let macrocell l be the neighboring macrocell(s) that uses the same pilot sequence as user k. The downlink signal to interference ratio (SIR) of user k when it connects to the MBS in the tagged macrocell, $\gamma_{k,0}$, is

$$\gamma_{k,0} = \beta_{k,0}^2 / \sum_{l \neq 0} \beta_{k,l}^2, \tag{5}$$

where $\beta_{k,0}$ is the factor accounting for the propagation loss and shadowing effects between the MBS and user k, and $\beta_{k,l}$ accounts for the propagation loss and shadowing factor between user k and the MBS in macrocell l. When neighboring macrocells use different values of τ, an MBS receives not only the pilot signals of users from other cells, but also uplink data signals from other cells. As analyzed in [19], the non-orthogonal uplink data signals also contaminate the channel estimation of other cells, and the resulting interference is a random variable bounded by the interference caused by pilot signals. Hence, we use (5) as a worst-case approximation in case the SIR cannot be measured by the MBS due to technical limits. When the values of τ are close to each other in different macrocells, such approximation would be highly reliable. Due to the mobility of users, we assume that $\gamma_{k,0}$ is updated with a period of T seconds.

The data rate of user k is given by [3]

$$R_{k,0} = \sum_{n=1}^{\alpha N} a_{k,n} \left(1 - \frac{T_p}{T_c}\tau\right)\left(\frac{T_u}{T_s}\right) \log\left(1 + \gamma_{k,0}\right), \tag{6}$$

where T_p is the time spent to transmit pilot for one user and $T_p = T_s$.[1] Due to channel reciprocity of the TDD mode, the CSI is acquired by the MBS using uplink pilots. Then, $\gamma_{k,0}$ and $R_{k,0}$ can be obtained by the MBS.

Similarly, let γ_j be the downlink SIR of WB between the MBS and SBS j, it is given by

$$\gamma_j = \beta_{j,0}^2 / \sum_{l \neq 0} \beta_{j,l}^2, \tag{7}$$

where $\beta_{j,0}$ is the factor accounts for the propagation loss and shadowing effects between the MBS and SBS j, and $\beta_{j,l}$ is the propagation loss and shadowing factor between SBS j and the MBS in macrocell l.

The data rate of the WB for SBS j is then given as

$$C_j = \sum_{n=1}^{\alpha N} b_{j,n} \left(1 - \frac{T_p}{T_c}\tau\right)\left(\frac{T_u}{T_s}\right) \log\left(1 + \gamma_j\right). \tag{8}$$

We assume that the time interval for uplink pilots of MUEs and WBs are used to send control information from SBS's to

[1] The value of T_p can also be optimized based on physical layer analysis. According to (6), a small value of T_p reduces the channel estimation overhead and increases $R_{k,0}$. However, the channel estimation quality may be degraded, resulting in decreased $R_{k,0}$. Due to space limit, we focus on frame level analysis and network scheduling problems, the potential of optimizing T_p with physical layer analysis can be investigated in future work.

SUEs, including CSI, power and channel schedule of SUEs. We also assume that equal resource allocation is applied to SUEs served by the same SBS so that proportional fairness can be achieved [15]. Let $\gamma_{k,j}$ be the average signal to noise plus interference ratio (SINR) of user k connecting to SBS j over a time period. The achievable data rate is given as

$$R_{k,j} = \left(1 - \frac{T_p}{T_c}\tau\right)\left(\frac{T_u}{T_s}\right)\frac{(1-\alpha)N}{\sum_{k=1}^{K} x_{k,j}} \log\left(1+\gamma_{k,j}\right). \quad (9)$$

We assume that the powers of SBS's and SUEs are adjusted to proper values so that the interference between different small cell users are controlled at an acceptable level. Unlike the MBS with massive MIMO, the effect of fast fading exists on the channel between an SUE and an SBS, resulting in frequently varying CSI. Therefore, it is infeasible to use the instantaneous CSI for scheduling purposes. To this end, $\gamma_{k,j}$ is based on the *time-averaged* CSI measured by the SBS over T seconds in the previous period, and it is updated every T seconds.

We aim to maximize the sum rate of a massive MIMO HetNet. Let $\mathbf{x}$, $\mathbf{a}$, and $\mathbf{b}$ denote the matrices of $\{x_{k,j}\}$, $\{a_{k,n}\}$, and $\{b_{j,n}\}$, respectively. The problem is formulated as

$$\begin{aligned}
\mathbf{P1}: \ & \max_{\{\mathbf{x},\mathbf{a},\mathbf{b},\tau\}} \textstyle\sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j}R_{k,j} && (10)\\
& \text{subject to: } \textstyle\sum_{j=0}^{J} x_{k,j} \le 1, \ k = 1,2,\ldots,K && (11)\\
& \textstyle\sum_{k=1}^{K} x_{k,j} \le S_j, \ j = 0,1,\ldots,J && (12)\\
& \textstyle\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \le \tau N_{\text{sm}}, \\
& \qquad n = 1,\ldots,\alpha N && (13)\\
& \textstyle\sum_{n=1}^{\alpha N} a_{k,n} \le E_k, \ k = 1,2,\ldots,K && (14)\\
& \textstyle\sum_{n=1}^{\alpha N} b_{j,n} \le F_j, \ j = 1,2,\ldots,J && (15)\\
& \textstyle\sum_{k=1}^{K} x_{k,j}R_{k,j} \le C_j, \ j = 1,2,\ldots,J && (16)\\
& \tau \le \tau_{\max}, \ \tau \in \mathcal{N}^{+} && (17)\\
& a_{k,n} \in \{0,1\}, \quad b_{j,n} \in \{0,1\}, \ x_{k,j} \in \{0,1\}, \\
& n = 1,\ldots,\alpha N, k = 1,\ldots,K, j = 0,\ldots,J. \\
& && (18)
\end{aligned}$$

In problem **P1**, constraint (11) is because each user can connect to at most one BS. We enforce an upper bound on the number of users that can be served by each BS in (12) to guarantee the QoS of users. Constraint (13) is directly from (4). By enforcing an upper bound on the number of channels that can be accessed by user k, constraint (14) is to guarantee fairness among the MUEs. Without such constraint, MUEs with high SIRs would be allocated with more channels than those with low SIRs, resulting in poor fairness performance.[2] Thus, the value of E_k for an MUE with high SIR is set to be lower than an MUE with low SIR.[3] Similarly, constraint (15) is to guarantee fairness among the WBs. Constraint (16) is due to the fact that the data rate of WB for SBS j should be larger than or equal to the sum rate of all SUEs served by SBS j. Constraint (17) enforces an upper bound for the number of symbols for pilot transmissions. Since both $\gamma_{k,0}$ and $\gamma_{k,j}$ are updated with the period of T, problem **P1** is also solved with the period of T.

III. Centralized Solution Algorithm

In this section, we develop a centralized iterative scheme to obtain the near optimal solution of **P1**. Problem **P1** is an integer programming problem with both coupling variables and coupling constraints, and constraint (16) is a nonlinear coupling constraint of two sets of variables. Thus, standard optimization techniques cannot be directly applied for the optimal solution.

To make the problem tractable, we decompose problem **P1** into (i) *WB and MUE resource allocation and pilot length optimization* problem and (ii) *user association* problem, and iteratively solve the two problems until convergence. At each iteration, we update the constraints of each problem to satisfy all constraints of the original problem.

A. Resource Allocation and Pilot Optimization

As can be seen in (6) and (9), $R_{k,0}$ is determined by $\mathbf{a}$; and $R_{k,j}$, $j = 1, ..., J$, is limited by $\mathbf{b}$. Due to constraint (16), the sum rate of all MUEs and WBs naturally serves as an upper bound for the sum rate of all users. Thus, it is reasonable to try to maximize this upper bound and iteratively tighten the gap, so that the final solution is a close approximation for the optimal solution of Problem **P1**. The problem of maximizing the sum rate of all MUEs and WBs for a given $\mathbf{x}$ is as follows.

$$\begin{aligned}
\mathbf{P2}: \ & \max_{\{\mathbf{a},\mathbf{b},\tau\}} \left(1 - \frac{T_p}{T_c}\tau\right) \cdot \\
& \left\{ \sum_{k=1}^{K}\sum_{n=1}^{\alpha N} a_{k,n}\log\left(1+\gamma_{k,0}\right) \right. \\
& \left. \qquad + \sum_{j=1}^{J}\sum_{n=1}^{\alpha N} b_{j,n}\log\left(1+\gamma_j\right) \right\} \\
& \text{subject to: } (13) - (18). \qquad (19)
\end{aligned}$$

[2] Due to the channel hardening effect, the channel gains across different frequencies are close to each other [31]. Thus, the dominant factor that impacts the performance of WBs and MUEs is the number of allocated channels. However, in other application scenarios where the channel response varies significantly over different frequencies, e.g., in a mm-wave network, frequency domain scheduling should be considered. Some existing approaches can be applied include proportional fairness scheduling [22] and bipartite matching based algorithm [40].

[3] The proper values of E_k and F_j depend on network topology, traffic pattern, and QoS requirement of users. In a specific system, E_k and F_j can be dynamically adjusted based on the QoS of users. When the data rate of a MUE or WB at the edge of cell is lower than a threshold, the values of E_k and F_j for the MUEs and WBs with highest data rates will be lowered in the next period. The adjustment strategy of E_k and F_j can be done with an offline training process for each cell.

Note that, constraint (16) can be written as $\sum_{n=1}^{\alpha N} b_{j,n} \geq \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)}$. Since $\sum_{n=1}^{\alpha N} b_{j,n}$ is always an integer, (16) is equivalent to $\sum_{n=1}^{\alpha N} b_{j,n} \geq \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$.

Suppose constraint (16) has already been satisfied for the WB of SBS j, then allocating more resources to this WB can not improve the actual sum rate of the users served by SBS j, while it potentially increases the value of τ, resulting in degraded system performance. Thus, (16) is an active constraint in problem **P2**. We have $\sum_{n=1}^{\alpha N} b_{j,n} = \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$. Combining this constraint with (15), we have

$$\sum_{n=1}^{\alpha N} b_{j,n} = \min\left\{ \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log\left(1+\gamma_j\right)} \right\rceil, F_j \right\} .j = 1, 2, \ldots, J. \tag{20}$$

To solve problem **P2**, we first relax the integer constraints by allowing them **a** and **b** to take any values in [0, 1], and τ to take any value in $[0, \tau_{\max}]$.

Lemma 1: The relaxed problem of **P2**, **P2-Relaxed**, *is a convex optimization problem.*

Proof: The objective function of **P2-Relaxed** is a sum of quadratic terms and linear functions, with the quadratic terms given as $-\tau a_{k,n}$ and $-\tau b_{j,n}$. It can be easily verified that the Hessian matrices of such quadratic terms are negative definite. Thus, the objective function is concave. Since all constraints are linear, **P2-Relaxed** is a convex optimization problem. ■

Since the decision variables are coupled in the constraints, we use a primal decomposition to transform problem **P2-Relaxed** into two levels of problems [26]. At the lower level, we find optimal solution of **a** and **b** for a given τ. Based on the solution of the lower level problem, the optimal value of τ is then obtained with a subgradient approach.

1) Optimal Solution of **a** *and* **b** *for Given* τ*:* With given τ, we have the following lower level problem of **P2-Relaxed**.

$$\begin{aligned} \mathbf{P3}: \ \max_{\{\mathbf{a},\mathbf{b}\}} & \sum_{k=1}^{K} \sum_{n=1}^{\alpha N} a_{k,n} \log\left(1+\gamma_{k,0}\right) \\ & + \sum_{j=1}^{J} \sum_{n=1}^{\alpha N} b_{j,n} \log\left(1+\gamma_j\right) \\ & \text{subject to: (13), (14), (18), and (20).} \end{aligned} \tag{21}$$

We can see that **P3** is a linear programming (LP), which can be solved with efficient methods such as simplex method. To analyze its property, we transform **P3** into the standard form by concatenating the columns of **a** and **b** alternately, given as

$$\begin{aligned} \tilde{\mathbf{y}} = \big[& a_{1,1}, \ldots, a_{K,1}, b_{1,1}, \ldots, b_{J,1}, a_{1,2}, \ldots, a_{K,2}, \\ & b_{1,2}, \ldots, b_{J,2}, \ldots, a_{1,\alpha N}, \ldots, a_{K,\alpha N}, b_{1,\alpha N}, \ldots, b_{J,\alpha N} \big]^T. \end{aligned} \tag{22}$$

Let **Z** be the constraint matrix corresponding to $\tilde{\mathbf{y}}$, as

$$\mathbf{Z} \doteq \begin{pmatrix} 1\,1\cdots1 & 0\,0\cdots0 & \cdots & 0\,0\cdots0 \\ 0\,0\cdots0 & 1\,1\cdots1 & \cdots & 0\,0\cdots0 \\ \vdots & \vdots & & \vdots \\ 0\,0\cdots0 & 0\,0\cdots0 & \cdots & 1\,1\cdots1 \\ 1\,0\cdots0 & 1\,0\cdots0 & \cdots & 1\,0\cdots0 \\ 0\,1\cdots0 & 0\,1\cdots0 & \cdots & 0\,1\cdots0 \\ \ddots & \ddots & \vdots & \ddots \\ 0\,0\cdots1 & 0\,0\cdots1 & \cdots & 0\,0\cdots1 \end{pmatrix} \tag{23}$$

The right hand side (RHS) of the LP is a $(\alpha N + J + K) \times 1$ vector, given by

$$\mathbf{d} = [\tau N_{\text{sm}}, \ldots, \tau N_{\text{sm}}, E_1, \ldots, E_K, \theta_1, \ldots, \theta_J]^T, \tag{24}$$

where $\theta_j = \min\left\{ \left\lceil \sum_{k=1}^{K} x_{k,j} R_{k,j} / \log\left(1+\gamma_j\right) \right\rceil, F_j \right\}$.

Lemma 2: The constraint matrix **Z** *is totally unimodular.*

Proof: Omitted due to lack of space, a similar case and proof can be found in [11]. ■

Property 1: If the constraint matrix of an LP satisfies totally unimodularity, and the RHS is integral, then it has all integral vertex solutions [24].

Property 2: If an LP has feasible optimal solutions, then at least one of the feasible optimal solutions occurs at a vertex of the polyhedron defined by its constraints [25].

Lemma 3: All the decision variables in the optimal solution to the relaxed LP, problem **P3***, are integers in* {0, 1}.

Proof: This lemma directly follows Lemma 2, Property 1, and Property 2. ■

2) Optimal Value of τ*:* Denote $g\left(\mathbf{a}(\tau), \mathbf{b}(\tau), \tau\right)$ and $f\left(\mathbf{a}(\tau), \mathbf{b}(\tau)\right)$ as the values of objective functions of **P2-Relaxed** and **P3** for a given τ, which are given in (19) and (21), respectively. Let $g^*(\tau)$ and $f^*(\tau)$ be their optimal values for a given τ, respectively. At the higher level of problem **P2-Relaxed**, we find the optimal value of τ by solving the following problem.

$$\mathbf{P4}: \ \max_{\{\tau\}} \ g^*(\tau). \tag{25}$$

Consider the objective function of **P2-Relaxed**, given as

$$g\left(\mathbf{a}(\tau), \mathbf{b}(\tau), \tau\right) = \left(1 - \frac{T_p}{T_c}\tau\right)\left(f\left(\mathbf{a}(\tau), \mathbf{b}(\tau)\right)\right). \tag{26}$$

Maximizing (26) is equivalent to maximizing the following

$$\log\left(1 - \frac{T_p}{T_c}\tau\right) + \log\left[f\left(\mathbf{a}(\tau), \mathbf{b}(\tau)\right)\right]. \tag{27}$$

Hence, problem **P4** is equivalent to the following problem

$$\begin{aligned} \max_{\{\tau\}} \ & \left\{ \log\left(1 - \frac{T_p}{T_c}\tau\right) + \log\left[f\left(\mathbf{a}^*(\tau), \mathbf{b}^*(\tau)\right)\right] \right\} \\ & \text{subject to: (17).} \end{aligned} \tag{28}$$

Let $h_1(\tau) = \log\left(1 - \frac{T_p}{T_c}\tau\right)$, $h_2(\tau) = \log\left[f\left(\mathbf{a}^*(\tau), \mathbf{b}^*(\tau)\right)\right]$, and $h(\tau) = h_1(\tau) + h_2(\tau)$. Since **P2-Relaxed** is a convex problem according to Lemma 1, we can apply primal decomposition to optimize $h_1(\tau)$ and

$h_2(\tau)$ separately [26]. It can be easily verified that $h_1(\tau)$ is a differentiable concave function. For any τ and τ', we have

$$\log\left(1-\frac{T_p}{T_c}\tau\right) \le \log\left(1-\frac{T_p}{T_c}\tau'\right) - \frac{T_p}{T_c - T_p\tau'}\left(\tau-\tau'\right).$$

Then, τ can be updated with the following gradient approach to maximize $h_1(\tau)$.

$$\tau^{[t+1]} = \tau^{[t]} - \frac{T_p}{T_c - T_p\tau^{[t]}}\rho^{[t]}, \tag{29}$$

where t is the index of iteration and $\rho^{[t]}$ is the step size.

To obtain the optimal solution of $h_2(\tau)$, we consider the following optimization problem

$$\begin{aligned}\mathbf{P5}:\ &\max_{\{\mathbf{a},\mathbf{b}\}} \log\left[f\left(\mathbf{a}(\tau),\mathbf{b}(\tau)\right)\right]\\ &\text{subject to: (13), (14), (18), and (20).}\end{aligned}$$

Lemma 4: Strong duality holds for problem **P5**.

Proof: Since problem **P5** is a convex problem, all the constraints are linear and the Slater condition reduces to feasibility [15], [23]. Thus strong duality holds. ■

Let λ_n^* be the optimal value of Lagrangian multiplier corresponding to the constraint $\sum_{k=1}^{K} a_{k,n}+\sum_{j=1}^{J} b_{j,n} \le \tau N_{sm}$. We consider the optimal solutions to **P5** for two different values, τ' and τ. Then, we have

$$\begin{aligned}
h_2\left(\tau'\right) &= \log\left[f\left(\mathbf{a}^*\left(\tau'\right),\mathbf{b}^*\left(\tau'\right)\right)\right]\\
&\overset{(a)}{=} \mathcal{L}\big(\mathbf{a}^*\left(\tau'\right),\mathbf{b}^*\left(\tau'\right),\boldsymbol{\lambda}^*\left(\tau'\right),\boldsymbol{\mu}^*\left(\tau'\right),\boldsymbol{\nu}^*\left(\tau'\right),\boldsymbol{\eta}^*\left(\tau'\right)\big)\\
&\overset{(b)}{\ge} \mathcal{L}\big(\mathbf{a}^*\left(\tau\right),\mathbf{b}^*\left(\tau\right),\boldsymbol{\lambda}^*\left(\tau'\right),\boldsymbol{\mu}^*\left(\tau'\right),\boldsymbol{\nu}^*\left(\tau'\right),\boldsymbol{\eta}^*\left(\tau'\right)\big)\\
&= \log\left[f\left(\mathbf{a}^*(\tau),\mathbf{b}^*(\tau)\right)\right]+\sum_{n=1}^{\alpha N}\lambda_n^*\left(\tau'\right)\left(\tau N_{sm}-\delta_n^*(\tau)\right)\\
&\qquad + \Phi + \sum_{n=1}^{\alpha N}\lambda_n^*\left(\tau'\right)\left(\tau' N_{sm}-\tau N_{sm}\right)\\
&\overset{(c)}{\ge} h_2(\tau) + N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^*\left(\tau'\right)\left(\tau'-\tau\right),
\end{aligned}\tag{30}$$

where $\delta_n^*(\tau) = \sum_{k=1}^{K} a_{k,n}^*(\tau)+\sum_{j=1}^{J} b_{j,n}^*(\tau)$, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, and $\boldsymbol{\eta}$ are the Lagrangian multipliers corresponding to other constraints. Φ is given as

$$\begin{aligned}
\Phi &= \sum_{k=1}^{K}\mu_k^*\left(\tau'\right)\left(E_k - \sum_{n=1}^{\alpha N} a_{k,n}(\tau)\right)\\
&\quad + \sum_{j=1}^{J}\nu_j^*\left(\tau'\right)\left(F_j - \sum_{n=1}^{\alpha N} b_{j,n}(\tau)\right)\\
&\quad + \sum_{j=1}^{J}\eta_j^*\left(\tau'\right)\left(\sum_{n=1}^{\alpha N} b_{j,n}(\tau) - \left\lceil\frac{R_{k,j}}{\log\left(1+\gamma_j\right)}\right\rceil\right).
\end{aligned}\tag{31}$$

In (30), equality (a) is due to strong duality, inequality (b) is due to the optimality of $\mathbf{a}^*\left(\tau'\right)$ and $\mathbf{b}^*\left(\tau'\right)$, and inequality (c) is due to the constraints of problem **P5** and the nonnegativity of all Lagrangian multipliers.

It follows (30) that

$$h_2(\tau) \le h_2\left(\tau'\right) + N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^*\left(\tau'\right)\left(\tau-\tau'\right). \tag{32}$$

By definition, $N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^*(\tau)$ is a subgradient of $h_2(\tau)$. The maximum value of $h_2(\tau)$ can be obtained by

$$\tau^{[t+1]} = \tau^{[t]} + N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^{*[t]}\rho^{[t]} \tag{33}$$

Lemma 5: Problem **P4** *can be solved by the following subgradient method.*

$$\tau^{[t+1]} = \tau^{[t]} + \left(N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^{*[t]} - \frac{T_p}{T_c - T_p\tau^{[t]}}\right)\rho^{[t]}. \tag{34}$$

Proof: According the principle of primal decomposition, $N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^{*[t]} - \frac{T_p}{T_c - T_p\tau^{[t]}}$ is a subgradient of $h(\tau)$, τ can be updated by combining (29) and (33). The optimal value of τ can be achieved until iteration converges. ■

There is a nice interpretation for (34). In each update, $N_{\rm sm}\sum_{n=1}^{\alpha N}\lambda_n^{*[t]}$ indicates the performance gain obtained by allocating more pilot symbols to WBs and MUEs, i.e., to increase τ. The second part, $\frac{T_p}{T_c - T_p\tau^{[t]}}$ indicates the performance loss caused by the reduced number of data symbols.

Denote $\eta^{[t]}$ as the subgradient of $h(\tau)$, $\eta^{[t]} = N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^{*[t]} - \frac{T_p}{T_c - T_p\tau^{[t]}}$, the convergence of the τ is shown in the following lemma.

Lemma 6: With step size set as $\rho^{[t]} = \frac{h(\tau^*)-h(\tau^{[t]})}{(\eta^{[t]})^2}$, *the sequence* $h(\tau^{[t]})$ *converges to its optimal value* $h(\tau^*)$ *with a speed faster than* $\{1/\sqrt{t}\}$ *as* $t\to\infty$.

Proof: Consider the optimality gap of τ, we have

$$\begin{aligned}
(\tau^{[t+1]}-\tau^*)^2 &\le \left(\tau^{[t]} + \frac{h(\tau^*)-h(\tau^{[t]})}{(\eta^{[t]})^2}\eta^{[t]} - \tau^*\right)^2\\
&= (\tau^{[t]}-\tau^*)^2 + \frac{\left(h(\tau^*)-h(\tau^{[t]})\right)^2}{(\eta^{[t]})^2}\\
&\quad +2(\tau^{[t]}-\tau^*)\eta^{[t]}\frac{h(\tau^*)-h(\tau^{[t]})}{(\eta^{[t]})^2}\\
&\le (\tau^{[t]}-\tau^*)^2 - \frac{\left(h(\tau^*)-h(\tau^{[t]})\right)^2}{(\eta^{[t]})^2}\\
&\le (\tau^{[t]}-\tau^*)^2 - \frac{\left(h(\tau^*)-h(\tau^{[t]})\right)^2}{\widehat{\eta}^2},
\end{aligned}$$

where $\widehat{\eta}$ is an upper bound of $|\eta^{[t]}|$. The first inequality is because $\tau^{[t+1]}$ should project to $[0,\tau_{\max}]$, the second inequality is due to the property of subgradient, given as $(\tau^{[t]}-\tau^*)\eta^{[t]} \le h(\tau^{[t]})-h(\tau^*)$. Summing the above inequality from $t=1$ to $t\to\infty$, we have

$$\sum_{t=1}^{\infty}\left(h(\tau^*)-h(\tau^{[t]})\right)^2 \le \widehat{\eta}^2(\tau^{[1]}-\tau^*)^2. \tag{35}$$

Suppose for contradiction, $\lim_{t\to\infty}\left(h(\tau^*)-h(\tau^{[t]})\right)\sqrt{t} > 0$. Then, there must be a sufficiently large t' and a positive

Algorithm 1 WB and MUE Resource Allocation and Pilot Length Optimization

```
Initialize $\tau$ ;
do
    Solve problem P5 to obtain $\lambda_n^*(\tau)$ ;
    Update $\tau$ with (34) ;
while ($\tau$ does not converge and $\tau \le \tau_{\max}$);
if $\tau < \tau_{\max}$ then
    Solve P3 with $\lfloor \tau^* \rfloor$ and $\lceil \tau^* \rceil$ to obtain $\mathbf{a}(\lceil \tau^* \rceil)$,
    $\mathbf{b}(\lceil \tau^* \rceil)$, $\mathbf{a}(\lfloor \tau^* \rfloor)$, and $\mathbf{b}(\lfloor \tau^* \rfloor)$ ;
    Use the results to compare the values of objective
    functions of P2. Then,
    $\tau^* = \arg\max_{\{\lfloor \tau^* \rfloor, \lceil \tau^* \rceil\}} \{g^*(\lfloor \tau^* \rfloor), g^*(\lceil \tau^* \rceil)\}$ ;
else
    Set $\tau^* = \tau_{\max}$ ;
end
Use $\tau^*$ to solve problem P3, and obtain the optimal a
and b ;
```

number ξ such that $\left(h(\tau^*) - h(\tau^{[t]})\right)\sqrt{t} > \xi, \forall t \ge t'$. Taking the square sum from t' to ∞, we have

$$\sum_{t=t'}^{\infty} \left(h(\tau^*) - h(\tau^{[t]})\right)^2 \ge \xi^2 \sum_{t=t'}^{\infty} \frac{1}{t} = \infty. \qquad (36)$$

It can seen that (36) contradicts (35). Thus, the hypothesis does not hold, we have

$$\lim_{t\to\infty} \frac{h(\tau^*) - h(\tau^{[t]})}{1/\sqrt{t}} = 0, \qquad (37)$$

this indicates that $h(\tau^{[t]})$ converges with a speed faster than that of $1/\sqrt{t}$. ■

Note that, the optimal τ to **P2-Relaxed** may not be an integer. Since **P2-Relaxed** is a convex problem, a simple way to find the optimal τ to **P2** is to compare the objective values of problem **P2** under $\lfloor \tau^* \rfloor$ and $\lceil \tau^* \rceil$, and select the larger one. As discussed in Lemma 3, the optimal solution to **P2-Relaxed** are integers for any given integer value of τ. Thus, such solution is also optimal to **P2**, we conclude that the optimal solution of **P2** can be obtained. The procedure of the proposed WB and MUE resource allocation and pilot length optimization scheme is summarized in Algorithm 1.

Lemma 7: The complexity of Algorithm 1 is upper bounded by $1/\varepsilon_1^2\varepsilon_2^2$, where ε_1 is the threshold of convergence for τ, ε_2 is the threshold of convergence for $\boldsymbol{\lambda}$.

Proof: According to Lemma 6 and (37), for a sufficiently large t and a sufficiently small ε_1, we have $h(\tau^*) - h(\tau^{[t]}) < 1/\sqrt{t}$. Thus, when $1/\sqrt{t} > \varepsilon_1$, $h(\tau^*) - h(\tau^{[t]})$ is guaranteed to be smaller than ε_1. Consequently, it takes less than $1/\varepsilon_1^2$ steps for the sequence $h(\tau^{[t]})$ to achieve an optimality gap that is less than ε_1, $t < 1/\varepsilon_1^2$. In the same way, the number of iterations for the convergence of $\boldsymbol{\lambda}$ is upper bounded by $1/\varepsilon_2^2$.

In Algorithm 1, each update of τ requires a set of optimal $\boldsymbol{\lambda}$. Thus, the total number of variable updates is upper bounded by $1/\varepsilon_1^2\varepsilon_2^2$, the complexity of Algorithm 1 is upper bounded by $1/\varepsilon_1^2\varepsilon_2^2$. ■

B. User Association under WB Constraints

For a given set of **a**, **b**, and τ, **P1** is reduced to the following user association problem.

$$\mathbf{P6}: \max_{\{\mathbf{x}\}} \sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j} R_{k,j}$$
$$\text{subject to: (11), (12), and (16)}$$
$$x_{k,j} \in \{0,1\}, \quad k = 1,2,\ldots,K, \quad j = 0,1,\ldots,J. \qquad (38)$$

Constraint (16) can be rewritten as

$$\sum_{k=1}^{K} x_{k,j}\left(\log\left(1+\gamma_{k,j}\right) - \frac{\sum_{n=1}^{\alpha N} b_{j,n}\log\left(1+\gamma_j\right)}{(1-\alpha)N}\right) \le 0, \quad j = 1,2,\ldots,J, \qquad (39)$$

which is a linear constraint on **x**.

To solve **P6**, we first relax the integer constraint of **x** by allowing all $x_{k,j}$ to take any value between $[0,1]$. Denote the relaxed problem as **P6-Relaxed**. The objective function of **P6-Relaxed** includes a weighted sum of $\frac{\sum_{k=1}^{K} x_{k,j}\log(1+\gamma_{k,j})}{\sum_{k=1}^{K} x_{k,j}}$, which is non-convex. Thus, only local optimal solution can be achieved with standard optimization techniques. However, if the values of $Q_j = \sum_{k=1}^{K} x_{k,j}$ are given, **P6-Relaxed** reduces to an LP.

Since $Q_j \le S_j$, the optimal solution of **P6-Relaxed** can be obtained by searching all possible combinations of $\mathbf{Q} = \{Q_1, ..., Q_J\}$ and solve the corresponding LPs. However, this results in a high complexity as $\prod_{j=1}^{J} S_j$ LPs need to be solved. We thus use this approach to obtain the initial optimal values of **Q** and update it with a more efficient approach. Recall that the system states are updated every T. Thus, in a low mobility environment, we can make use of **Q** in the previous period as an approximation to the **Q** of the current period. Then, $\{R_{k,j}\}$ becomes independent of **x**, given as

$$R_{k,j} = \frac{1}{Q_j}\left(1 - \frac{T_p}{T_c}\tau\right)\left(\frac{T_u}{T_s}\right)(1-\alpha)N\log\left(1+\gamma_{k,j}\right).$$

P6-Relaxed is thus transformed to the following LP.

$$\mathbf{P7}: \max_{\{\mathbf{x}\}} \sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j} R_{k,j}$$
$$\text{subject to: (11), (12), and (39)}$$
$$x_{k,j} \in [0,1], \quad k = 1,2,\ldots,K, \quad j = 0,1,\ldots,J. \qquad (40)$$

Since **P7** is an LP, the cutting plane method [27] can be applied to obtain its optimal *integer solution*, and such solution is also optimal to **P6** for a given **Q**. As users may dynamically join or leave the network, the traffic load of each BS varies over time, the approximation of **Q** might be inaccurate. However, a key observation is that *load balancing* can be achieved by solving **P7**. When Q_j is larger than its optimal value, $R_{k,j}$ would be small. Then fewer users would be connected to SBS j after the update with the solution of **P7**, resulting in a decreased Q_j. Thus, the value of Q_j is expected to stay close to its optimal value, and the solution is expected to be near-optimal.

In case the user distribution drastically changes and handover frequently happen (e.g., during rush hours), which can be detected by each BS when measuring the CSI of nearby users, $\mathbf{Q}$ should be updated by solving **P6-Relaxed** with searching over all $\mathbf{Q}$. Due to its high complexity, such update is carried out at a timescale much larger than T.

C. Iterative Scheme with Near-Optimal Solution

In this section, we propose an iterative approach to obtain the near-optimal solution of the original problem by solving the *WB and MUE resource allocation and pilot length optimization* problem and the *user association* problem iteratively until convergence. The iterative scheme is a three-stage process to guarantee that all constraints are satisfied as well as minimizing the gap of the two problems. The proposed three-stage process is based on the following facts.

Lemma 8: Under optimal user association solutions, given fixed values of Q_j of other BS's, the sum rate of all users served by SBS j decreases as Q_j increases.

Proof: According to (9), $\sum_{k=1}^{K} R_{k,j}$ is proportional to $\frac{\sum_{k=1}^{K} x_{k,j}\log(1+\gamma_{k,j})}{\sum_{k=1}^{K} x_{k,j}}$, which can be interpreted as the average spectral efficiency of users served by SBS j.

Consider an optimal user association with a given feasible set of $\mathbf{Q}$. To maximize the sum rate, the users served by SBS j must be the first Q_j users with the highest spectral efficiencies, i.e., the highest SINRs. Thus, when the values of Q_j for other BS's are fixed, the average spectral efficiency of users served by SBS j decreases as Q_j increases. ■

Property 3: In **most cases***, the users served by SBS j are the first Q_j users with highest SINRs, and the sum rate of all users served by SBS j decreases as Q_j increases.*

Compared to Lemma 8, we remove the assumption that the values of Q_j for other BS's are fixed. The only exception of Property 3 happens when a user k' originally served by a neighboring SBS j' is handed over to SBS j due to an increase of $Q_{j'}$, while the SINR of this user is higher than at least one of the users currently served by SBS j. Suppose user k has a lower SINR than user k' when served by SBS j. Then both users are likely to be cell-edge users, and the coverage areas of SBS j and SBS j' are likely to overlap. Hence, the exception case happens when both Q_j' and Q_j increase and a cell-edge user is handed over to SBS j. As a result, when the SBS's are not densely deployed, the exception case would not happen.

a) Stage I: In the first stage, we aim to guarantee that constraints (12), $\sum_{k=1}^{K} x_{k,j} \leq S_j$, are always satisfied for all SBS's. We begin with solving the initial MUE and WB resource allocation and pilot length optimization problem without considering the constraint on WB data rate, $\sum_{n=1}^{\alpha N} b_{j,n} = \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$. This corresponds to the case of setting the initial values of the RHS of (20) to be F_j. Let **P8** be the LP generated by removing constraints (12) from **P7**, **P8** can be solved by the same approach as **P7**. Then, we find the optimal user association under WB constraints by solving **P8**. With such initial solution, C_j may be low for SBS j, $\sum_{k=1}^{K} R_{k,j}$ is bounded by a low value. As in Property 3, a large number of users are expected to be assigned to SBS j to achieve a low value of $\sum_{k=1}^{K} R_{k,j}$, which may violate constraint (12) and be infeasible to **P7**. Thus, we first solve **P8** to find the set of SBS's that violate the WB constraint, and then enforce additional constraints to **P8** for feasibility.

With the solution of **P8**, if constraint (12) of SBS j is not satisfied, **P8** is updated by adding constraint $\sum_{k=1}^{K} x_{k,j} = S_j$. Then, we update $R_{k,j}$ by keeping the first S_j highest SINR users to be served by SBS j. After that, we update constraint (20) for SBS j with the updated $x_{k,j}$ and $R_{k,j}$. This way, both constraints for SBS j are satisfied; the WB resource allocation and user association for SBS j become feasible. Based on Property 3, by keeping the first S_j highest SINR users, the value of $\sum_{k=1}^{K} R_{k,j}$ is expected to be the largest under a feasible and optimal solution of **P7**. This results in the smallest change on the RHS of constraint (20) for SBS j. Thus, the change of the polyhedron defined by $\mathbf{Z}$ is minimized, resulting in a smallest reduction of the objective function. Then, we solve the MUE and WB resource allocation and pilot length optimization problem with the updated constraint (20) for SBS j. After that, we use the solution to solve **P8** in the next iteration. Such process is repeated until all constraints (12) are satisfied for all SBS's. After the process is converged, we enter the second stage.

b) Stage II: In the second stage, we aim to minimize the performance gap between the two problems, so that $C_j - \sum_{k=1}^{K} x_{k,j} R_{k,j}$ is minimized. The motivation of minimizing such gap is because allocating more channels to WBs leads to increased value of τ and decreased data rates of all users, it is desirable that the data rate provided by each WB is sufficiently utilized by each SBS. To minimize the gap at each SBS, we find the SBS's with $\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$, and update these constraints as

$$\sum_{n=1}^{\alpha N} b_{j,n} = \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log\left(1+\gamma_j\right)} \right\rceil. \tag{41}$$

Then, we obtain the optimal $\{\mathbf{a}, \mathbf{b}, \tau\}$ with the updated constraints as in Section III-A. With $\{\mathbf{a}, \mathbf{b}, \tau\}$, we solve **P7** to obtain the optimal $\mathbf{x}$. Such process is repeated until $\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$ does not hold for any SBS.

c) Stage III: In the third stage, we aim to guarantee that the WB constraints of all SBS's are satisfied after the updates in the second stage. With the update in the second stage, the values of $\sum_{n=1}^{\alpha N} b_{j,n}$ are reduced, which may cause an increased ratio of $\sum_{n=1}^{\alpha N} a_{k,n} / \sum_{n=1}^{\alpha N} b_{j,n}$ for some users. Hence, under the optimal solution of **P8**, these users may switch to the MBS. According to Property 3, the sum rate of SBS's that served these users in the previous iteration are expected to increase, resulting violation of the WB constraints. To deal with this situation, we can adjust and update the values of $\sum_{n=1}^{\alpha N} b_{j,n}$ with (20), and we repeat this process until the WB constraints of all SBS's are satisfied.

The procedure of the proposed iterative scheme is summarized in Algorithm 2.

Algorithm 2 Iterative Scheme to Obtain a Near-Optimal Solution to Problem **P1**

```
Initialize Set the RHS of (20) as $F_j$ ;
do
    Obtain $\{\mathbf{a}, \mathbf{b}, \tau\}$ with Algorithm 1;
    Solve P8 to obtain $\mathbf{x}$ ;
    for $j = 1 : J$ do
        if $(\sum_{k=1}^{K} x_{k,j} > S_j)$ then
            Set $\sum_{k=1}^{K} x_{k,j} = S_j$ ;
            Update (20) for SBS $j$ ;
            Add constraint $\sum_{k=1}^{K} x_{k,j} = S_j$ to P8 ;
        end
    end
    Solve P8 to obtain updated $\mathbf{x}$ ;
    Obtain $\{\mathbf{a}, \mathbf{b}, \tau\}$ with updated $\mathbf{x}$ using Algorithm 1 ;
while $(\sum_{k=1}^{K} x_{k,j} \leq S_j$ does not hold for all $j$ );
do
    for $j = 1 : J$ do
        if $(\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil)$ then
            Update $\sum_{n=1}^{\alpha N} b_{j,n}$ with (20) ;
        end
    end
    Update $\{\mathbf{a}, \mathbf{b}, \tau\}$ with Algorithm 1 ;
    Update $\mathbf{x}$ by solving P8 ;
while $(\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$ holds for any $j$ );
do
    for $j = 1 : J$ do
        if $(\sum_{n=1}^{\alpha N} b_{j,n} \leq \left\lfloor \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rfloor)$ then
            Update $\sum_{n=1}^{\alpha N} b_{j,n}$ with (20) ;
        end
    end
    Update $\{\mathbf{a}, \mathbf{b}, \tau\}$ with Algorithm 1 ;
    Update $\mathbf{x}$ by solving P8 ;
while $(\sum_{n=1}^{\alpha N} b_{j,n} \leq \left\lfloor \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rfloor$ holds for any $j$ );
```

D. Remarks on Practical Concerns

1) Quasi-Static Channel Between MBS and SBS's: Due to the fixed locations of SBS's, the channels between SBS's and MBS are quasi-static [7]. As a result, the CSI of WBs can be updated less frequently compared to that of MUEs. This property can be employed to enhance the system performance. In most time periods, the SBS's can use some channels for WB transmission without sending pilots on these channels, resulting in reduced pilot length. Thus, we can assign the WBs to use all the αN channels to increase the data rate. In such scenario, the problem formulation can be derived from Problem **P1** with modifications on the constraints.

Due to the different frequencies of CSI update for MUE and WB, there are two cases at different time periods.

- **First case**: Both WBs and MUEs need to send pilots. When the CSI of WBs needs to be updated, all WBs are allocated with one pilot sequence on each channel so that the CSI of WBs on all channels can be obtained. This corresponds to set $b_{j,n} = 1$ for $j = 1, \ldots, J$, $n = 1, \ldots, \alpha N$. In addition, the constraint $\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j$ can be removed from Problem **P1** since **b** is given.
- **Second case**: Only the MUEs need to send pilots. In these periods, the MBS uses the CSI obtained in the first case until the next update of CSI of WB. Then, the constraint $\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\text{sm}}$ in Problem **P1** should be modified to $\sum_{k=1}^{K} a_{k,n} \leq \tau N_{\text{sm}}$. Since the WBs are allocated with all channels, we have $b_{j,n} = 1$ for $j = 1, \ldots, J$, $n = 1, \ldots, \alpha N$. Same as the first case, the constraint $\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j$ is also removed.

With given τ, it can be easily verified that the constraint matrix of the linear programming for solving $\{a_{k,j}\}$ is unimodular for both cases. Thus, we can obtain the optimal $\{\mathbf{a}, \tau\}$ using the same approach as in Algorithm 1. Then, we apply Algorithm 2 to obtain the solutions for both cases.

2) A Combination of Wired and Wireless Backhaul: In case of dense SBS deployment with heavy traffic load, a long pilot length (i.e., large τ) is required, resulting in degraded system performance. To mitigate such bottleneck as well as preserving the benefits of wireless backhaul, a combination of wired and wireless backhaul is desirable. With proper configuration, a good tradeoff between performance and cost can be achieved.

With a combination of wired and wireless backhaul, the problem formulation needs to be adjusted accordingly. We assume that the data rate of wired backhaul is sufficiently high so that the constraint $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$ can always be satisfied. Let Ω be the set of SBS's that use wireless backhaul, then the constraints $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$ and $\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j$ only apply to $j \in \Omega$. The constraints (20) and (39), which are derived from $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$, are also applied to $j \in \Omega$ only. For the constraint $\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\text{sm}}$, we replace the term $\sum_{j=1}^{J} b_{j,n}$ to $\sum_{j \in \Omega} b_{j,n}$. The solution under such new scenario can be obtained with the same approach in Algorithm 2 with the updated constraints (20) and (39). Specifically, since the SBS's with wired backhaul have no impact on the pilot optimization, we still maximize the sum rate of wireless backhaul and MUE by solving Problem **P2** with Algorithm 1. For user association, the constraint $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$ does not apply to the SBS's with wired backhaul, and the problem can be solved with the same approach presented before.

IV. Distributed Solution Scheme

In the centralized scheme, global information is required for centralized control, which usually leads to better performance. But acquiring the global information may incur considerable overhead, which may be infeasible in a large scale network. In this section, we propose a distributed scheme by formulating a noncooperative repeated game among all users. In the repeated game, each user distributively makes its own decision. We demonstrate that the game will converge to an NE.

Algorithm 3 Distributed User Association Strategy for BS j

```
while (convergence not achieved) do
    if (BS $j$ holds more than $S_j$ proposals) then
        Put the top $S_j$ users with the highest SINRs in
        the waiting list and reject the other users ;
    else
        Put all users in the waiting list ;
    end
end
```

A. Distributed User Association

We formulate a repeated game among all users, the strategy of each user is to decide its serving BS. Due to the tradeoff in MUE and WB resource allocation, we set a price for using one channel such that the number of channels used by MUEs and WBs can be controlled at proper values. The utility of user k is defined as

$$\begin{cases} \mathcal{U}_{k,0} = \omega_k \log\left(R_{k,0}\right) - p \cdot \sum_{n=1}^{\alpha N} a_{k,n} \\ \mathcal{U}_{k,j} = \omega_k \log\left(R_{k,j}\right) - p \cdot \dfrac{\sum_{n=1}^{\alpha N} b_{j,n}}{\sum_{k=1}^{K} x_{k,j}}, \quad j = 0, \ldots, J. \end{cases} \tag{42}$$

where ω_k is the evaluation of user k for data rate and p is the price of using one channel. When user k is served by an SBS, the cost of channels for the WB is shared by all users that are served by the SBS. In (42), $\sum_{n=1}^{\alpha N} a_{k,n}$ is set by each user to be a fixed value that maximizes its utility, given as $\sum_{n=1}^{\alpha N} a_{k,n} = \arg\max_{\left\{\sum_{n=1}^{\alpha N} a_{k,n}\right\}} \left\{\mathcal{U}_{k,0}\right\} = \omega_k / p$. For $\sum_{n=1}^{\alpha N} b_{j,n}$, it is a variable given by (41), which is affected by other users' decisions. The strategy of each user is

$$x_{k,j^*} = 1, \quad j^* = \arg\max_j \{\mathcal{U}_{k,j}\}. \tag{43}$$

To maximize the sum rate under constraint $\sum_{k=1}^{K} x_{k,j} = S_j$, it is reasonable to assume that each BS serves the top S_j users with highest SINRs. The user association strategy of BS's is summarized in Algorithm 3.

Each user has a *preference list* for all BS's, the order of the list is determined by the order of $\mathcal{U}_{k,j}$, e.g., the BS with the largest $\mathcal{U}_{k,j}$ is the first in the preference list of user k. Since Q_j is unknown before the repeated game, the initial preference list of each user is determined by values of SINRs when connecting to different BS's. The proposed repeated game has the following two stages.

In the *first* stage, each user proposes to the top BS in its preference list. Then, BS's respond to the proposals according to Algorithm 3.

In the *second* stage, each BS j broadcasts the value of Q_j to all users. Then each user k updates its preference list with $R_{k,j}$. A user proposes to another BS under the following cases.

Case 1: The proposal of the user is rejected.

Case 2: A higher utility can be achieved by switching to another BS j' and one of the two conditions is satisfied: (i) $Q_{j'} < S_{j'}$, (ii) $Q_{j'} = S_{j'}$, and there is a user k' currently in the waiting list of BS j' such that $R_{k,j'} > R_{k',j'}$.

If user k is rejected by BS j, it marks BS j as *unavailable* in its preference list. Then, users in these two cases propose to the top BS among remaining available BS's. Once receiving the proposals, each BS compares the new proposals with those in its waiting list, and makes decisions according to Algorithm 3. If a user switches from BS j to BS j' as described in Case 1, the users that once marked BS j as *unavailable* change the status of BS j to *available*. Given the BS decisions, each user then updates its preference list and makes another round of proposal if one of the two cases is satisfied. The repeated game is continued until convergence of user association is achieved.

After convergence, the MBS replaces constraint (14) with $\sum_{n=1}^{\alpha N} a_{k,n} = \omega_k / p$ and update constraint (15) with (20). It then determines $\{\mathbf{a}, \mathbf{b}, \tau\}$ as in Section III-A.

B. Convergence Analysis

The convergence property of the repeated game is given in Theorem 1, which shows that an NE can be achieved.

Theorem 1: The repeated game converges to a Nash equilibrium that is optimal for each user.

Proof: Suppose the game does not converge. Then, there must be a user k that is currently served by BS j who wishes to propose to another BS j'. Obviously, Case 1 does not hold since user k is served by BS j. Then, Case 2 holds, there is another BS j' such that $\mathcal{U}_{k,j'} > \mathcal{U}_{k,j}$ and BS j' is marked as *available* by user k. If condition (i) is satisfied, $Q_{j'} < S_{j'}$, then user k would have already switched to BS j', which contradicts to the fact that it is served by BS j. If condition (ii) is satisfied, $Q_{j'} = S_{j'}$, then there must another user k' that is served by BS j' such that $R_{k,j'} > R_{k',j'}$, i.e., BS j' prefers user k over user k'. Since user k' is in the waiting list of BS j' while user k is not, it must be the case that user k has never proposed to BS j' before. However, since $\mathcal{U}_{k,j'} > \mathcal{U}_{k,j}$, user k must have proposed to BS j' before BS j, which is also a contradiction. Thus, the repeated game converges.

From the above analysis, we can see that the utility of each user cannot be further improved given the strategies of other users. Thus, the strategy of each user is the *best response* to the strategies of other users when the repeated game converges. We conclude that the repeated game converges to an NE. ■

The order of users that start the proposed process affects the system performance, as different NEs would be achieved. Such randomness results in performance loss of distributed scheme compared to the centralized one.

V. Simulation Study

We validate the proposed centralized and distributed schemes with MATLAB simulations. The scenario is based on a cellular system with hexagonal macrocells, and we consider the sum rate of all users in a tagged macrocell area. The MBS is located at the center, the SBS's and users are randomly distributed in the macrocell area. The radius of a macrocell is 500 m. The slow fading factor, $\beta_{k,0}$, is based on the ITU path loss model [28] and a lognormal shadowing with standard deviation of 10 dB. The coherence bandwidth is 150 kHz. We use the parameters of downlink LTE symbol for each OFDM symbol. The spacing between subcarriers is 15 kHz, then

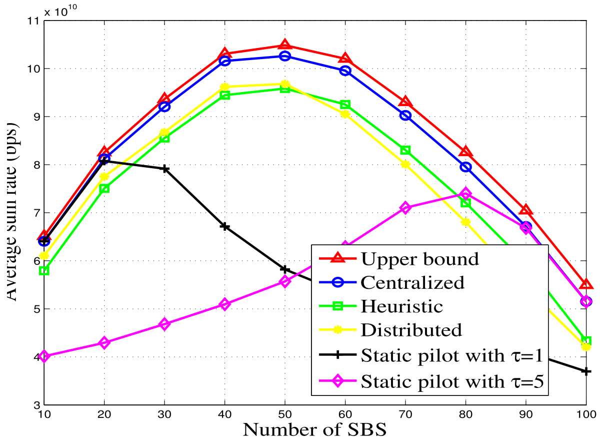

Fig. 1. Average sum rates of different schemes versus the number of SBS (200 users).

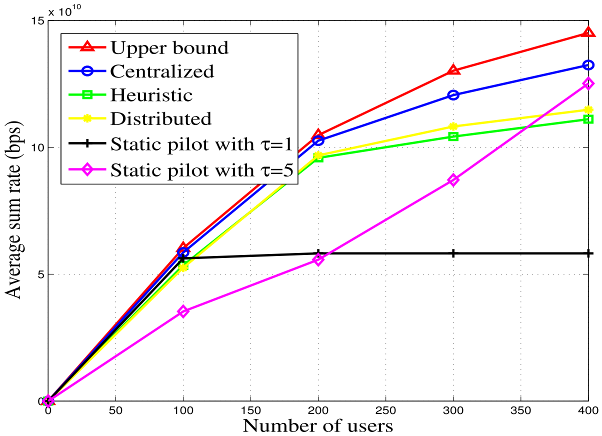

Fig. 2. Average sum rates of different schemes versus the number of users (20 SBS's).

$N_{\text{sm}} = 10$; the useful symbol duration $T_u = 1/\Delta_f = 66.7$ ms; and $T_s = T_p = 72$ ms. The coherence time is $T_c = 720$ ms, so each frame has 10 OFDM symbols, and we set $\tau_{\max} = 5$. The total bandwidth is 4 MHz, so the total number of channels is 40. We assume $\alpha = \frac{1}{2}$; then 20 channels are allocated to MUEs and WBs and the other 20 channels are allocated to SUEs. The powers of SBS's are set according to the iterative water-filling scheme [29], with an upper bound of 30 dBm. The upper bounds of $\sum_{k=1}^{K} x_{k,j}$ are set to be $S_j = 20$ for SBS's and $S_0 = 50$ for MBS, respectively.

We compare the proposed schemes with a heuristic scheme, termed *Heuristic*, for user association. Heuristic is based on Property 3 and is derived by making a modification on the centralized scheme. Specifically, instead of solving **P8** at each iteration of the centralized scheme, the set of users served by each SBS is determined with a greedy approach. In each round, we select the user with highest SINR to be served by SBS j and update the value of $\sum_{k=1}^{K} R_{k,j}$. We continue such process until $\sum_{k=1}^{K} R_{k,j} \leq C_j$ is satisfied. We also consider the case based on [15], in which pilot length is not considered for optimization and τ is set as a fixed value (termed *Static pilot*). For Static pilot, the solution of $\{\mathbf{a}, \mathbf{b}, \tau\}$ is based on the solution procedure in Section III-A. For Heuristic, we apply the same procedure of the proposed centralized scheme except the user association strategy. Since the performance of the distributed scheme depends on the value of p, we set p to the value that achieves the maximal sum rate. We also consider the value of the objective function of problem **P2** under optimal solution as an upper bound for comparison.

The sum rate performances of different schemes are presented in Figs. 1 and 2. In Fig. 1, it can be seen that the performances of all schemes first increase and then decrease as the number of SBS's grows. This is because a larger τ is required as the number of SBS's increases, and the interference between neighboring small cells degrades the average SINRs of SUEs. Both the centralized and distributed schemes outperform Static pilot, demonstrating that a performance gain can be achieved with dynamically adjusted τ. The performance of the centralized scheme is close to its upper bound, since we iteratively minimize the performance gap of two problems in the second stage of the iterative scheme given in Algorithm 2.

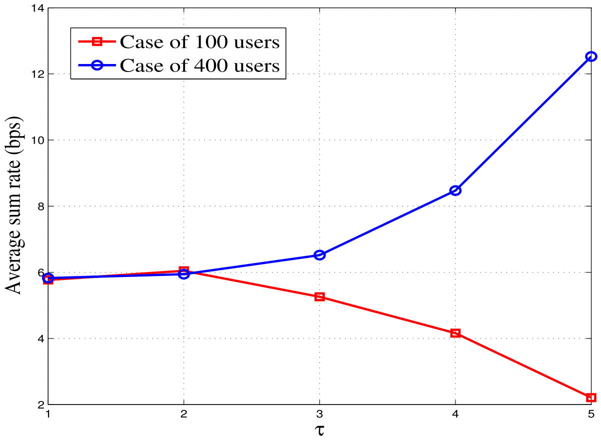

Fig. 3. Average sum rates versus the value of τ under 2 different numbers of users (20 SBS's).

It is also observed that the performance of Heuristic is close to that of the centralized scheme when the number of SBS's is small, due to the fact that Property 3 is more reliable when SBS's are not close to each other, and a user would have a significant difference in data rates by connecting to different SBS's. The distributed scheme also achieves a satisfactory performance since users are charged for using channels, resulting in efficient resource utilization. For Static pilot, the case of $\tau = 1$ achieves better performance than the case of $\tau = 5$ when the number of SBS's is small, since a small τ can accommodate the requirements of all WBs. However, when the number of SBS's is large, a larger τ provides better performance since the increased demand for WB data rates can be satisfied.

Fig. 2 shows the performances under different numbers of users, where similar trends can be observed. When the number of users increases, the sum rate of users with $\tau = 1$ remains constant. This is because the resources for MUEs and WBs are quite limited. As a result, a considerable proportion of users cannot be served by any BS.

In Fig. 3, the performance of Static pilot with different τ values is evaluated. When τ is large, the performance with 100 users is significantly worse than that with 400 users; however, when τ is small, the performance with 400 users

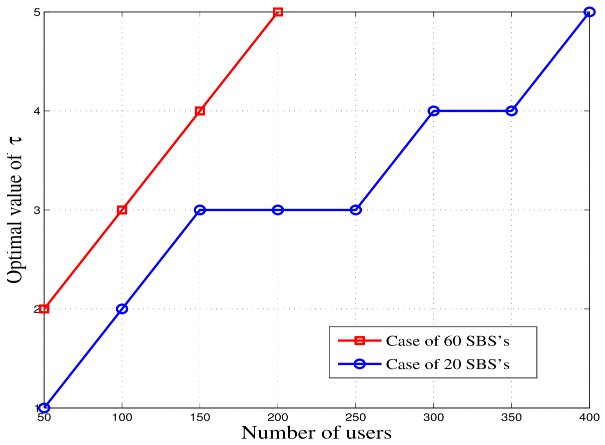

Fig. 4. Optimal value of τ under different numbers of SBS's (200 users).

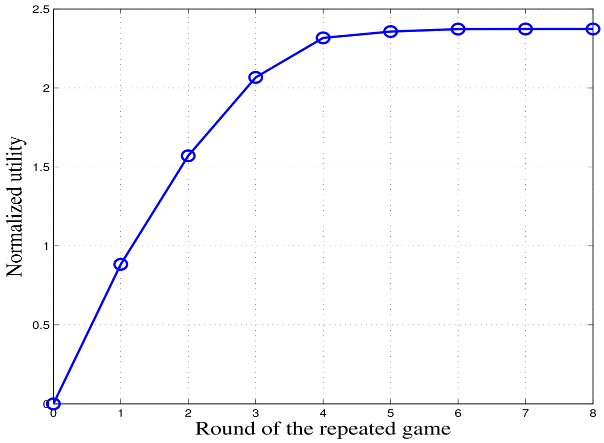

Fig. 5. Convergence of the repeated bidding game (200 users and 20 SBS's).

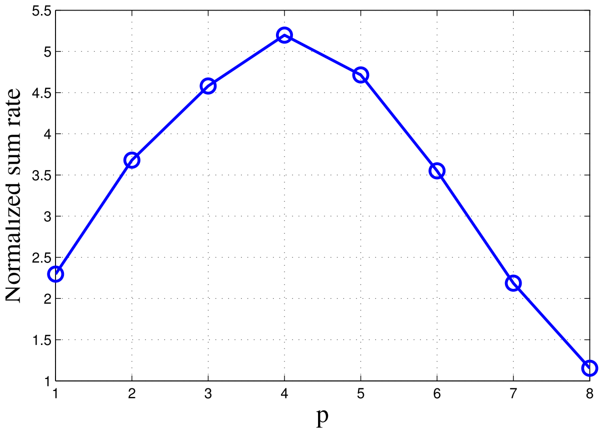

Fig. 6. Normalized sum rate versus the value of p (200 users and 20 SBS's).

becomes worse than that with 100 users. This shows that a small value of τ significantly limit the system performance in case of larger number of users, and τ needs to be dynamically adjusted to prevent considerable performance loss of Static pilot under different traffic patterns. The optimal values of τ under different numbers of SBS's and users are presented in Fig. 4. The optimal τ increases with both the number of SBS's and the number of users, since the more resources are required to satisfy the increasing demand.

An example of the repeated game is given in Fig. 5. We can see that the game converges after several rounds and a maximum sum utility is achieved upon convergence.

We also present an example to evaluate the impact of price p on the system performance in Fig. 6. By setting p to a proper value, each user makes rational decision on channel usage, the value of τ can be set to a proper value.

VI. Related Work

The Massive MIMO technology has attracted lots of interests in recent years. A comprehensive introduction on the fundamentals of signal processing issues can be found in [30]. An overview and analysis for upper layer techniques in massive MIMO systems is presented in [6]. The potentials, limits and possible research problems of massive MIMO were presented in [31]. As an important application scenario, massive MIMO HetNet has also been widely studied. The key technical aspects of massive MIMO HetNet include user association [10], [20] and interference management [7], [34]. The RTDD architecture for massive MIMO HetNet was first considered in [7]. With RTDD, the interference occurs between MBS and SBS's. Then, the MBS can employ zero-forcing beamforming based on the estimated channel covariance beween MBS and SBS's. Compared to these works, we consider wireless connection between the MBS and each SBS, and propose a cross-layer optimization framework.

The HetNet with WB has been studied in several prior works. Since another type of transmission is added over transmissions between users and BS's, interference management becomes a key issue under certain system assumptions and has been investigated in [35] and [36]. In [37], an load-aware design on spatial multiplexing was proposed to improve the energy efficiency of a HetNet with WB. A recent overview on resource management of 5G HetNet with WB was presented in [14]. In this paper, we integrate massive MIMO with WB and deal with the challenges with an adaptive frame design.

User association in HetNet is another closely related issue. A recent survey on user association of 5G network can be found in [38]. In [39], a near-optimal user association scheme was proposed, and the proposed scheme can be implemented distributively with a dual decomposition. To deal with the integer constraint, several approximation algorithms were proposed in [40] with the objective of minimizing the maximum load among all BS's. In [41], user association was considered from the perspective of maximizing the utilities of users, and different spectrum allocation strategies were jointly considered. In [42], a traffic-aware dynamic user association was considered through the cooperation of SBS's. Due to the special network architecture of a massive MIMO HetNet with WB, we optimize the network performance with joint frame design, resource allocation, and user association in this paper.

VII. Conclusions

In this paper, we considered the problem of joint frame design, resource allocation, and user association to maximize the sum rate of a massive MIMO HetNet with WBs. We formulated a nonlinear integer programming problem and proposed a centralized iterative scheme to obtain a near-optimal solution. We also proposed a distributed scheme by formulating a repeated game among all users and prove that the game converges to an NE. Simulation results show that the proposed schemes outperform several benchmark schemes.

REFERENCES

[1] M. Feng and S. Mao, "Adaptive pilot design for massive MIMO HetNets with wireless backhaul," in *Proc. IEEE SECON*, San Diego, CA, USA, Jun. 2017, pp. 1–9.

[2] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[5] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access J.*, vol. 2, pp. 947–959, Aug. 2014.

[6] M. Feng and S. Mao, "Harvest the potential of massive MIMO with multi-layer techniques," *IEEE Netw.*, vol. 30, no. 5, pp. 40–45, Sep./Oct. 2016.

[7] K. Hosseini, J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO and small cells: How to densify heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 5442–5447.

[8] E. Björnson, M. Kountouris, and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," in *Proc. Int. Conf. Telecommun. (ICT)*, Casablanca, Morocco, May 2013, pp. 1–5.

[9] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, Mar. 2016.

[10] Y. Xu and S. Mao, "User association in massive MIMO HetNets," *IEEE Syst. J.*, vol. 11, no. 1, pp. 7–19, Mar. 2017.

[11] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station on-off switching strategy for energy efficient massive MIMO HetNets," in *Proc. INFOCOM*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[12] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov. 2014.

[13] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: Challenges and solution approaches," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 22–31, Oct. 2015.

[14] N. Wang, E. Hossain, and V. K. Bhargava, "Backhauling 5G small cells: A radio resource management perspective," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 41–49, Oct. 2015.

[15] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3251–3268, May 2016.

[16] B. Li, D. Zhu, and P. Liang, "Small cell in-band wireless backhaul in massive MIMO systems: A cooperation of next-generation techniques," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 7057–7069, Dec. 2015.

[17] H. Tabassum, A. H. Sakr, and E. Hossain, "Analysis of massive MIMO-enabled downlink wireless backhauling for full-duplex small cells," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2354–2369, Jun. 2016.

[18] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, Oct. 2015.

[19] F. Fernandes, A. Ashikhmin, and T. L. Marzetta, "Inter-cell interference in noncooperative TDD large scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 192–201, Feb. 2013.

[20] Q. Ye, O. Y. Bursalioglu, H. C. Papadopoulos, C. Caramanis, and J. G. Andrews, "User association and interference management in massive MIMO HetNets," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2049–2065, May 2016.

[21] E. Bjӧrnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral effciency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[22] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 210–212, Mar. 2005.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] A. Schrijver, *Theory Linear Integer Programming*. Hoboken, NJ, USA: Wiley, Jun. 1998.

[25] C. Berenstein and R. Gay, *Complex Variables: An Introduction*. New York, NY, USA: Springer, 1997.

[26] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[27] R. Gomory, "Outline of an algorithm for integer solutions to linear programs," *Bull. Amer. Math. Soc.*, vol. 64, no. 5, pp. 275–278, Sep. 1958.

[28] *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000*, document Rec. ITU-R M.1225, 1997.

[29] W. Yu, "Multiuser water-flling in the presence of crosstalk," in *Proc. IEEE ITA Workshop*, San Diego, CA, USA, Jan. 2007, pp. 414–420.

[30] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[31] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[32] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[33] K. Zheng, L. Zhao, J. Mei, B. Shao, W. Xiang, and L. Hanzo, "Survey of large-scale MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1738–1760, 3rd Quart., 2015.

[34] A. Adhikary, H. S. Dhillon, and G. Caire, "Massive-MIMO meets HetNet: Interference coordination through spatial blanking," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1171–1186, Jun. 2015.

[35] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-Aho, "Backhaul-aware interference management in the uplink of wireless small cell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5813–5825, Nov. 2013.

[36] L. Sanguinetti, A. Moustakas, and M. Debbah, "Interference management in 5G reverse TDD HetNets with wireless backhaul: A large system analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1187–1200, Jun. 2015.

[37] H. H. Yang, G. Geraci, and T. Q. S. Quek, "Energy-efficient design of MIMO heterogeneous networks with wireless backhaul," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4914–4927, Jul. 2016.

[38] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.

[39] 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 Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[40] H. Zhou, S. Mao, and P. Agrawal, "Approximation algorithms for cell association and scheduling in femtocell networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 3, pp. 432–443, Sep. 2015.

[41] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: A utility perspective," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1025–1039, Jun. 2015.

[42] M. Feng, T. Jiang, D. Chen, and S. Mao, "Cooperative small cell networks: High capacity for hotspots with interference mitigation," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 108–116, Dec. 2014.

Mingjie Feng (S'15) received the B.E. and M.E. degrees in electrical engineering from the Huazhong University of Science and Technology in 2010 and 2013, respectively. He was a visiting student at the Department of Computer Science, Hong Kong University of Science and Technology, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA. His research interests include cognitive radio networks, heterogeneous networks, massive MIMO, mm-wave network, and full-duplex communication. He was a recipient of a Woltosz Fellowship at Auburn University.

Shiwen Mao (S'99–M'04–SM'09) received the Ph.D. degree in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA, in 2004. He is the Samuel Ginn Distinguished Professor and Director of the Wireless Engineering Research and Education Center, Auburn University, Auburn, AL, USA. His research interests include wireless networks and multimedia communications. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society. He received the 2015 IEEE ComSoc TC-CSR Distinguished Service Award, the 2013 IEEE ComSoc MMTC Outstanding Leadership Award, and the NSF CAREER Award in 2010. He was a co-recipient of the Best Demo Award from the IEEE SECON 2017, the Best Paper Awards from the IEEE GLOBECOM 2016 and 2015, the IEEE WCNC 2015, the IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is on the Editorial Boards of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE INTERNET OF THINGS JOURNAL, the IEEE MULTIMEDIA, and the ACM GETMOBILE.

Tao Jiang (M'06–SM'10) received the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004. He is currently the Chair Professor with the School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan, China. He has authored or co-authored over 200 technical papers in major journals and conferences and nine books/chapters in the areas of communications and networks. He was a recipient of the NSFC Distinguished Young Scholars Award in 2013. He was recognized as among the Most Cited Chinese Researchers announced by Elsevier in 2014, 2015, and 2016. He has served or is serving as an Associate Editor of some technical journals in communications, including the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE INTERNET OF THINGS JOURNAL. He is the Associate Editor-in-Chief of CHINA COMMUNICATIONS.