Distributed User Association and Resource Allocation Algorithms for Three Tier HetNets

Swaroop Gopalam, Stephen V. Hanly and Philip Whiting
School of Engineering, Macquarie University, Sydney, Australia
swaroop.gopalam@hdr.mq.edu.au,{stephen.hanly,philip.whiting}@mq.edu.au

***Abstract*—In this paper, we consider joint optimization of user association and resource allocation in three tier HetNets. We formulate the objective of minimizing the resources required to clear a given set of files, as a linear program. We show that the optimal user association is determined by a rate-biasing rule, where a bias value is associated with each BS. We show that each rate-bias value crucially only takes values from a finite set which we characterize. We present a complete analytical solution along with new structural results. Using these results, we present efficient distributed algorithms for optimal control of three tier HetNets. The method involves a 1D search for a resource variable at the macro-level, and 2D search at the pico-level for a resource variable and a bias value. We apply our results to a variety of hierarchical network examples.**

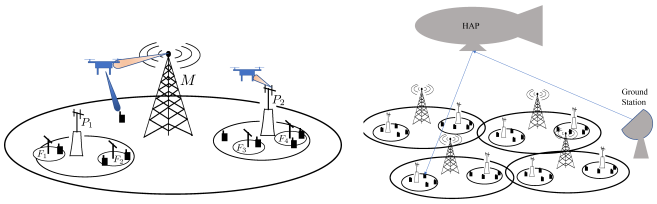
I. Introduction

Heterogeneous Networks (HetNets) consist of low power base stations (BSs) such as pico cells and femto cells deployed to operate in same region as the traditional macro cellular infrastructure [1]. These small cells increase the capacity due to better spatial re-use of spectrum. Future 5G networks are expected to be even more heterogeneous with wireless access to user equipments (UEs) simultaneously available via multiple technologies, including new technologies such as mmWave and aerial BSs. As demand increases and cells get smaller, there will be an increased number of tiers in future HetNet architectures.

Stochastic geometry based approaches traditionally employed for studying HetNets provide analytical results on coverage and SINR distributions, but are not suitable for real-time control. The optimization based literature has focused on cell association and resource allocation for two tier HetNets, and to date, there is no complete analytical solution for HetNets with more than two tiers. With the increased complexity of 5G networks, there is a need for studying the problem in general cases with more than two tiers. This paper provides a complete solution to the joint three tier resource allocation & cell association problem. Results in this paper allow for a wide range of new and emerging wireless networks to be analyzed within a common framework. Examples of these networks include 1) mmWave small cell networks, 2) networks with aerial platforms and 3) multi-tier radio access technologies. In the following paragraphs, we discuss important future technologies which can be treated under this framework.

This research was supported in part by the Australian Research Council under grant DP180103550. It was also supported by a iMQRTP PhD scholarship from Macquarie University.

5G mmWave cellular networks are expected to have very small cell sizes (which leads to increase in tiers in 5G HetNet architectures) due to high path-loss and blocking experienced at mmWave frequencies [1], [2]. Connectivity of mmWave links can be highly intermittent due to blocking by mobile objects. Cell association schemes are needed to offload (and provide continual service to) the blocked mmWave UEs [2]. Wireless backhaul solutions are being investigated to enable dense deployments of mmWave cells [1]. In section V, we optimize HetNets with mmWave small cells, using our framework. We consider the effect of wireless backhaul and blocking, and provide insights into design of handover schemes and backhaul planning.



(a) Network with Drones (b) HAP Network

Fig. 1: HetNet with UAVs

In addition to terrestrial networks, wireless communication using aerial platforms is also being considered for future networks [3]. In low altitude platform applications, unmanned aerial vehicles (UAVs) are deployed as mobile BSs to provide wireless access, or as UEs requiring access from the existing BSs [4]. In high altitude platform (HAP) applications, aircrafts or airships are deployed at altitudes of 17 to 22 km in the stratosphere to provide wireless connectivity over a large area [5]. HAPs have a very large coverage area, typically a few macro-sites, adding an extra tier at the top of the existing terrestrial network. These networks can be modelled as three tier HetNets as shown in Fig. 1.

Our framework has the following features which are common to all the above mentioned applications. 1) The BSs can be divided into tiers based on their coverage area. Generally, the higher tier cells have BSs at higher altitudes which cover larger areas. Several smaller cells can operate in the coverage area of a high tier cell. 2) A UE can potentially associate and get service from multiple BSs in different tiers, and 3) A higher tier BS may cause debilitating interference to the smaller cells in its coverage area, and resource partitioning can be used to address this. From two to three tiers, there

is an increase in the dimensionality of the joint optimization problem, e.g., two resource variables per UE to three per UE. Therefore, complexity of algorithms is a crucial consideration, which we address in the paper.

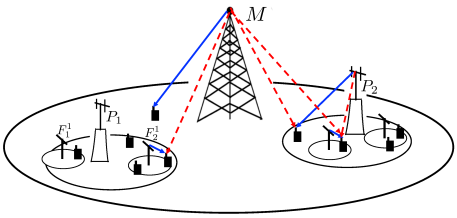


Fig. 2: A general three tier HetNet. The blue lines depict the BS to UE links and the red lines depict the interference

Joint user association and resource allocation problems were studied for HetNet control in several works in the literature. In [6]–[10], the approach was utility maximization. In [11]–[13], stochastic geometry was used to derive results. In [14]–[20], optimization for flow-based models was considered. [21]–[24] considered utility maximization including power control.

Although some works modelled k-tier HetNets (for $k \geq 3$), they had drawbacks. In [6], [9], [16], [20] resource partitioning between tiers to avoid cross-tier interference was not considered. The solutions in [12], [13], [17] are not adaptive to the changes in traffic, and hence not suitable for real-time control. Also in [12], [13], [17], the same bias value was applied to all the BSs in a tier, which is restrictive. Centralized solutions were proposed in [10], [21], [22], [24]. Only heuristic solutions were given in [21], [23], [24].

In this paper, we consider the objective of clearing a given set of files in the network using minimum possible resources, and refer to it as the *minimum time clearing problem*. In our prior work [18], [19], [25], similar formulations were used to derive joint optimization results for two tier HetNets. The three tier problem was considered in an early investigation in [26]. However, the proposed solution involved a high dimensional search (equal to number of femto BSs) to find the solution, which makes it prohibitive for real-time implementation.

We consider the problem of jointly optimizing user association and resource allocation in a three tier downlink HetNet. We refer to the tiers as macro, pico and femto tiers. We formulate the minimum time clearing problem as a linear program (LP). We show that by fixing the time allocated to small cells, the LP can be decomposed into several (equal to the number of pico BSs) smaller independent LPs. It follows that significant parallelization can be achieved by solving these LPs simultaneously at the corresponding pico BSs. We then show that the user association is determined by a set of rate-bias multipliers, one multiplier per BS. The problem of finding the multipliers using conventional approaches leads to a high dimensional search e.g., [6], [26]. In contrast, we present new structural results which enable us to propose more efficient algorithms with reduced complexity. The contributions of the paper are given in the following

- We provide a tractable framework for joint-optimization of user association and resource allocation in three tier HetNets. Our framework allows for full partitioning of resources between tiers, critical when there is strong cross-tier interference. The framework can be applied in a real-time manner for optimal control, or can be used as an offline tool for downlink capacity analysis.
- We present distributed algorithms to find the optimal solution under the proposed framework. The algorithms are highly efficient due to the new structural results we obtain in the paper.
- We show that each rate-bias multiplier (corresponding to a BS) crucially only takes values from a small discrete set. For a pico P_i, the size of the set is less than $0.5(|U_i|^2 + |U_i|)$, where $|U_i|$ is the number of UEs covered by the pico P_i.
- We further show that the solution of the LP at each pico BS is determined by just two parameters: the femto time allocation and the pico rate-bias multiplier.

We now present the outline of the paper. In Section II, we describe the system model and problem formulation. In Section III, we present the main results of the paper. In Section IV, we present the numerical results derived using simulations. In Section V, we treat the HetNet with mmWave small cells and backhaul under the framework developed in the paper.

II. System Model and Problem Formulation

We consider a 3-tier HetNet as shown in Fig. 2, with time division for resource partitioning. There are N pico BSs labelled as $\{P_i\}_{i=1}^{N}$, operating in the coverage area of the macro BS M. There are N_i femto BSs operating in the coverage area of a pico P_i, labelled as $\{F_j^i\}_{j=1}^{N_i}$. We consider significant cross-tier interference in our model (as it is the case in HetNets, e.g., [12]). Therefore, two different tier BSs with overlapping coverage areas are not allowed to transmit at the same time, i.e., no two BSs in $\{M, P_i, F_j^i\}$ can transmit at the same time. For user association, each UE can associate with at most 3 BSs - the macro M, a pico P_i and a femto F_j^i, (where j, i depend on the UE location). Let U_j^i denote the set of UEs that can associate with the femto F_j^i, and $U_i := \bigcup_{j=1}^{N_i} U_j^i$ denote the set of UEs that can associate with the pico P_i[1]. Let $U := \bigcup_{i=1}^{N} U_i$ denote the set of all the UEs.

Let $\mathcal{B}_j^i$ denote the set of all the BSs excluding F_j^i, P_i and M. The rate (in bits/sec) of the link between the femto F_j^i and a UE $u \in U_j^i$ (provided the BSs M and P_i are muted) is given by

$$T_u = B \log_2(1 + p_{F_j^i} g_{F_j^i,u} / (\sigma^2 + I_{\mathcal{B}_j^i,u}))$$

where B is the transmission bandwidth, $g_{b,u}$ is channel gain between the BS b and the UE u. $g_{b,u}$ includes the antenna gain, path loss and shadowing loss. p_b is the transmit power of BS b and σ^2 is the noise floor. The term $I_{\mathcal{B}_j^i,u}$ is the aggregate interference caused to the transmissions from F_j^i

[1]For notational simplicity, we do not explicitly model the UEs that have no femto connectivity and only have coverage from a pico P_i and macro M. Such UEs can be treated as being in range of a virtual femto $F_{N_i+1}^i$ which provides zero rate. For these UEs, some of the thresholds calculated in the paper will be infinite, but this only means that the UEs do not associate with the virtual femto that provides zero rate.

to u by the BSs in $\mathcal{B}_j^i$. We treat $I_{\mathcal{B}_j^i,u}$ as constant interference which depends on $\mathcal{B}_j^i$ and u, i.e., as another noise source[2]. Similar assumptions are commonly adopted in the literature, e.g., [16], [20].

Similarly, let R_u (and S_u) denote the rate (in bits/sec) of the link between $u \in U_j^i$ and the macro M (and the pico P_i respectively), provided the other two BSs in $\{M, P_i, F_j^i\}$ are muted. The BS and UE are assumed to obtain the rate information using pilot signal measurements.

A. Minimum time clearing LP

Under the model, a UE u has to download a file of size D_u bits. The file can be downloaded from any BS that is in range. A UE $u \in U_j^i$ can possibly download a part of the file from each of BSs M, P_i & F_j^i. For $u \in U_j^i$, let x_u (y_u & z_u) denote the amount of file (in bits) downloaded from the macro M (pico P_i & femto F_j^i resp.). For this setup, the minimum time (in sec) required to clear the traffic of all the UEs is formulated as LP (1-4).

$$\min_{x_u,y_y,z_u,\pi,\epsilon_i \geq 0} \quad \pi + \sum_{u \in U} x_u/R_u \tag{1}$$

$$\text{s.t.} \sum_{u \in U_i} y_u/S_u \leq \pi - \epsilon_i, \ \forall i \in \{1, 2, \ldots, N\} \tag{2}$$

$$\sum_{u \in U_j^i} z_u/T_u \leq \epsilon_i, \ \forall j \in \{1, \ldots, N_i\}, i \in \{1, \ldots, N\} \tag{3}$$

$$x_u + y_u + z_u = D_u, \ \forall u \in U \tag{4}$$

In (1), $\sum_{u \in U} x_u/R_u$ is the time used by the macro and π is the total time used by the small cell BSs. Out of time π, $\pi - \epsilon_i$ is used by a pico P_i (see (2)) and the rest, ϵ_i is used by each of the femtos $\{F_j^i\}_{j=1}^{N_i}$ (see (3)). Note that the BSs M, P_i and F_j^i are allocated different times under the LP, thus avoiding cross-tier interference.

Fixing a value of π, LP (1-4) can be decomposed into N independent LPs, one for each pico P_i. The LP involving P_i is formulated as LP (5). Let $f_i(\pi)$ denote the optimal solution of LP (5) for a given π. The solution of LP (1-4) is given by $\min_{\pi \in [0,\infty)} \pi + \sum_{i=1}^N f_i(\pi)$.

$$\begin{aligned} & \min_{x_u,y_u,z_u,\epsilon_i \geq 0} \sum_{u \in U_i} x_u/R_u \quad \text{s.t.} \\ \alpha \text{ constraint}: \quad & \sum_{u \in U_i} y_u/S_u \leq \pi - \epsilon_i \\ \beta_j \text{ constraint}: \quad & \sum_{u \in U_j^i} z_u/T_u \leq \epsilon_i, \ \forall j \in \{1, \ldots, N_i\} \\ \gamma_u \text{ constraint}: \quad & x_u + y_u + z_u = D_u, \ \forall u \in U_i \end{aligned} \tag{5}$$

Note that $f_i(\pi)$ can be computed at pico P_i. Hence, the computation of clearing time $\pi + \sum_{i=1}^N f_i(\pi)$ can be parallelized and distributed over the picos. Fig 4(a) (on page 6) depicts a distributed scheme using message passing for such a computation. Since the function $\pi + \sum_{i=1}^N f_i(\pi)$ is convex, the optimal π and clearing time can be found by methods such as a line search. This is the strategy that we will follow to find the solution of LP (1-4).

In the rest of the paper, we will focus on finding $f_i(\pi)$ at a pico P_i for an arbitrary $i \in \{1, \ldots N\}$, i.e., we will focus on the solution of LP (5). The notation used in the paper is summarised in the following table.

Notation	Description	
x_u; y_u; z_u	Amount of file (in bits) allocated to $u \in U_j^i$ by macro M; pico P_i; femto F_j^i resp.	
R_u; S_u; T_u	Rate of the link (in bits/sec) between $u \in U_j^i$ and macro M; pico P_i; femto F_j^i resp.	
$\pi - \epsilon_i$	Time allocated for transmissions of pico P_i	
ϵ_i	Time allocated for simultaneous transmissions of femtos $\{F_j^i\}_{j=1}^{N_i}$	
1; α; β_j	Rate bias multiplier corresponding to macro M; pico P_i; femto F_j^i resp.	α, β_j and ρ_u^α are defined in Theorem 1 below
R_u; $\frac{S_u}{\alpha}$; $\frac{T_u}{\beta_j}$	Biased rate of $u \in U_j^i$ from macro M; pico P_i; femto F_j^i resp.	
D_u	Total file size of u (in bits)	
ρ_u^α	$\min\{T_u/R_u, \alpha T_u/S_u\}$	

III. Main Results

For any $u \in U_i$, let x_u^*, y_u^* and z_u^* denote the value of x_u, y_u and z_u respectively under an optimal solution of LP (5). Let $\boldsymbol{x} := [x_u]_{u \in U_i}$, $\boldsymbol{y} := [y_u]_{u \in U_i}$ and $\boldsymbol{z} := [z_u]_{u \in U_i}$. We present the main results as the following Theorems. For proofs, refer to Appendix A.

Theorem 1 (Rate biasing rule)**.** *There exist optimal rate-bias multipliers*[3]*, $\alpha^* > 0$ corresponding to P_i and $\beta_j^* > 0$ corresponding to F_j^i, $j \in \{1, \ldots N_i\}$ which determine the user-association as follows. For any $u \in U_j^i$,*
1) $x_u^* > 0$, only if $R_u = 1/\gamma_u^*$
2) $y_u^* > 0$, only if $S_u/\alpha^* = 1/\gamma_u^*$
3) $z_u^* > 0$, only if $T_u/\beta_j^* = 1/\gamma_u^*$.
where $1/\gamma_u^* := \max\{R_u, S_u/\alpha^*, T_u/\beta_j^*\}$ is the maximum biased rate. Further, define $\rho_u^\alpha := \min\{\alpha T_u/S_u, T_u/R_u\}$. For any $u \in U_j^i$, the statements of the flow chart in Fig 3 hold true.

Theorem 1 states that the optimal user association is determined by a rate-biasing rule. The macro rate is not biased (or equivalently, the bias is 1). For a UE $u \in U_j^i$, the pico and femto biased rates are S_u/α^* and T_u/β_j^* respectively. A UE u associates with the BS (or BSs) providing the highest biased rate $1/\gamma_u^* = \max\{R_u, S_u/\alpha^*, T_u/\beta_j^*\}$. e.g., 1) If $R_u > \max\{S_u/\alpha^*, T_u/\beta_j^*\}$, the UE u associates with macro M (Macro UE case in Fig. 3), 2) If $R_u = S_u/\alpha^* > T_u/\beta_j^*$, the UE u associates with both the macro M and pico P_i (Macro-Pico split UE case in Fig. 3).

Fig. 3 presents all the possible cases of allocation that can occur under the rate-biasing rule. Theorem 1 provides a partial characterization of the solution of LP (5) via Fig. 3. Given the optimal multipliers, Theorem 1 determines the allocation for non-split UEs (as given in Fig. 3). But, the split UEs may receive a part of the file from each associated BS which is not

[2]Strictly speaking, interference depends on the set of BSs in $\mathcal{B}_j^i$ transmitting in a given slot. However, this is upper-bounded by the worst-case interference $\sum_{b \in \mathcal{B}_j^i} p_b g_{b,u}$. In this paper, we take $I_{\mathcal{B}_j^i,u}$ to be the worst-case value to be definite.

[3]The rate bias multipliers α, β_j and the variable γ_u correspond to the dual-variables of LP (5). For complete details, refer to Appendix B and Appendix E.

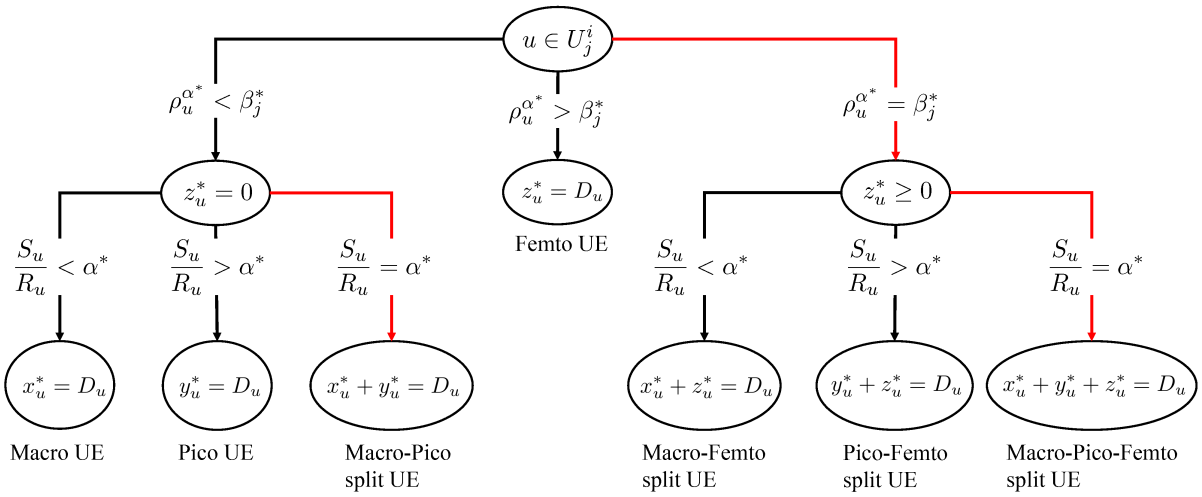


Fig. 3: User Association flow chart. The conditions leading to split UE cases are colored in red

given here. Algorithm 1 (on page 4) provides the full solution. To develop the algorithm, we first present the following two theorems concerning the optimal solution and the multipliers.

Theorem 2 (Finite set of rate-bias multipliers)**.** *Let* $\rho_u^\alpha := \min\{\alpha T_u/S_u, T_u/R_u\}$. *The optimal rate bias multipliers* α^* *and* β_j^* *take values from finite sets* A *and* B_j *respectively, as follows*

(i) $\alpha^* \in A$, *where*

$$A := \bigcup_{j=1}^{N_i} \{\frac{S_a T_b}{T_a R_b} : \frac{T_b}{R_b} \leq \frac{T_a}{R_a}, \frac{T_a}{S_a} \leq \frac{T_b}{S_b}\}_{(a,b)\in U_j^i \times U_j^i}$$

(ii) $\beta_j^* \in B_j := \{\rho_u^\alpha : u \in U_j^i\}_{\alpha \in A},\ j \in \{1, \dots N_i\}$

Note that $|A| \leq \sum_{j=1}^{N_i} \frac{|U_j^i|^2 + |U_j^i|}{2}$.

Given that the rate bias multipliers lie in a finite set (characterized in Theorem 2), it is tempting to implement a discrete search to find the optimal rate bias multipliers. However as explained in the previous paragraph concerning Theorem 1, such knowledge does not provide the allocation for split UEs. The following theorem forms the basis of our algorithmic solution, which achieves two goals, 1) it provides the full allocation, including split UEs, and 2) it reduces the dimensionality of the problem to just 2.

Theorem 3 (Allocation function)**.** *There exists an allocation function* $\Theta : \mathbb{R}_+ \times [0, \pi] \to \mathbb{R}_+^{3|U_i|+N_i}$ *(defined in Algorithm 1), which provides a mapping from a pair of* $\{\alpha, \epsilon_i\} \in \mathbb{R}_+ \times [0, \pi]$ *to a solution* $[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}]$ *of LP (5) and the femto rate bias multipliers* $\boldsymbol{\beta} = [\beta_j]_{j=1}^{N_i} > 0$ *as follows.*

$$[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}] = \Theta(\alpha, \epsilon_i)$$

Moreover, the function satisfies $[\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{z}^*, \boldsymbol{\beta}^*] = \Theta(\alpha^*, \epsilon_i^*)$, *where* α^* *is the optimal pico rate-bias multiplier and* ϵ_i^* *is the optimal femto time in LP (5).*

Theorem 3 provides the complete solution of LP (5), given by $\Theta(\alpha^*, \epsilon_i^*)$. It also shows that the solution is determined by just two variables - α^* and ϵ_i^*. Hence, a search over 2 parameters: discrete search for α^* over A and a continuous search for ϵ_i^* over $[0, \pi]$, can be implemented to solve LP (5) (in contrast to a high dimensional search for N_i+1 multipliers, e.g., [6], [26]).

Algorithm 1 Allocation Function $\Theta(\alpha, \epsilon_i)$

1: Run Algorithm 2 to obtain $\{\mathcal{F}_j(\alpha, \epsilon_i)\}_{j=1}^{N_i}$ and to evaluate $\boldsymbol{\beta}, \boldsymbol{z}, a, b, \delta$. // Femto allocation step // A special case that can occur is when two split UEs, a femto-pico split UE a and a femto-macro split UE b are in U_j^i for some j (See step 5 in Algorithm 2). In this case, z_a, z_b will be determined by Algorithm 3 in the next step. Here, $\delta = z_a/T_a + z_b/T_b$ is the available femto-time to be shared between a and b.
2: Run Algorithm 3 to obtain $\mathcal{P}(\alpha, \epsilon_i, \boldsymbol{z}, a, b, \delta)$, and to evaluate $\boldsymbol{y}, z_a, z_b$. // Pico allocation step
3: $x_u = D_u - y_u - z_u,\ \forall u \in U_i$. // Macro allocation step
4: **return** $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}$

The allocation function $\Theta(\alpha, \epsilon_i)$ of Theorem 3 (given in Algorithm 1) is defined using the the femto allocation functions $\mathcal{F}_j(\alpha, \epsilon_i), j \in \{1, \dots N_i\}$ given in Algorithm 2 and a pico allocation function $\mathcal{P}([\mathcal{F}_j(\alpha, \epsilon_i)]_{j=1}^{N_i})$ given in Algorithm 3.

The femto allocation function $\mathcal{F}_j(\alpha, \epsilon_i)$ determines the femto multiplier β_j and femto allocation $[z_u]_{u \in U_j^i}$ for the femto F_j^i (in step 1 of Algorithm 1). The pico allocation function $\mathcal{P}([\mathcal{F}_j(\alpha, \epsilon_i)]_{j=1}^{N_i})$ takes the outputs from the femto functions and determines the pico allocation $[y_u]_{u \in U_i}$ for the pico P_i (in step 2 of Algorithm 1). The macro allocation $[x_u]_{u \in U_i}$ can be completed by step 3 in Algorithm 1. The individual steps in Algorithm 2 and Algorithm 3 are justified by the Lemmas (contained in the Appendices) mentioned in the corresponding steps.

Algorithm 2 Femto Allocation Algorithm $\mathcal{F}_j(\alpha, \epsilon_i)$

1: Initialize $flag_j = 0$ // This flag is used to note the occurrence of two split users case, and 0 by default.
2: Sort $u \in U_j^i$ in descending order of ρ_u^α such that $\rho_{u_1}^\alpha \geq \dots \geq \rho_{u_K}^\alpha$. // where $K = |U_j^i|$

3: **if** $\sum_{k=1}^{K} D_{u_k}/T_{u_k} \leq \epsilon_i$ **then** // No split user case.

$z_{u_k} = D_{u_k}$ for $1 \leq k \leq K$ (See Lemma 4 in Appendix C)

$\beta_j = \rho^{\alpha}_{u_K}$ (See Lemma 10 in Appendix E)

4: **else if** $\exists l \leq K$ such that $\rho^{\alpha}_{u_l} \neq \rho^{\alpha}_{u_k}, \forall k \neq l$ and $\sum_{k=1}^{l-1} D_{u_k}/T_{u_k} \leq \epsilon_i < \sum_{k=1}^{l} D_{u_k}/T_{u_k}$ **then** // One femto split user case

$$z_{u_k} = \begin{cases} D_{u_k} & \text{for } 1 \leq k \leq l-1 \\ T_{u_l}(\epsilon_i - \sum_{k'=1}^{l-1} D_{u_{k'}}/T_{u_{k'}}) & \text{for } k = l \\ 0 & \text{for } l+1 \leq k \leq K \end{cases}$$

$\beta_j = \rho^{\alpha}_{u_l}$ (See Lemma 5 in Appendix C)

5: **else if** $\exists l \leq K-1$ such that $\rho^{\alpha}_{u_l} = \rho^{\alpha}_{u_{l+1}}$ and $\sum_{k=1}^{l-1} D_{u_k}/T_{u_k} \leq \epsilon_i < \sum_{k=1}^{l+1} D_{u_k}/T_{u_k}$ **then** // Two femto split users case. We now set a flag to denote that this case occurred.

6: $flag_j = 1$, $a := u_l$, $b := u_{l+1}$ // femto-pico split UE is a, femto-macro split UE is b

$$z_{u_k} = \begin{cases} D_{u_k} & \text{for } 1 \leq k \leq l-1 \\ 0 & \text{for } l+2 \leq k \leq K \end{cases}$$

$\beta_j = \rho^{\alpha}_{u_l} = \rho^{\alpha}_{u_{l+1}}$ (See Lemma 6 in Appendix C)

7: $\delta := \epsilon_i - \sum_{k=1}^{l-1} D_{u_k}/T_{u_k}$. // δ is the femto time for a and b. Step 7 of Algorithm 3 determines z_a & z_b.

8: **end if**

9: **return** $\{z_u\}_{u \in U^i_j - \{a,b\}}, \beta_j, a, b, \delta$

Algorithm 3 Pico Allocation Algorithm $\mathcal{P}(\alpha, \epsilon_i, \boldsymbol{z}, a, b, \delta)$

1: $W :=$ the set of UEs $u \in U_i - \{a, b\}$ such that $D'_u := D_u - z_u > 0$ // not femto only UEs

2: Sort $w_k \in W$ in descending order such that $S_{w_1}/R_{w_1} > \ldots > S_{w_{|W|}}/R_{w_{|W|}}$.

3: **if** $flag_j = 0, \forall j \in \{1, \ldots N_i\}$ **then** // Two split users case did not occur in $\{\mathcal{F}_j(\alpha, \epsilon_i)\}_{j=1}^{N_i}$

4: Find $l \leq |W|$ such that $\sum_{k=1}^{l-1} D'_{w_k}/S_{w_k} < \pi - \epsilon_i \leq \sum_{k=1}^{l} D'_{w_k}/S_{w_k}$

$$y_{w_k} := \begin{cases} D'_{w_k} & \text{for } 1 \leq k \leq l-1 \\ 0 & \text{for } l+1 \leq k \leq |W| \end{cases}$$

$$y_{w_l} := S_{w_l}(\pi - \epsilon_i - \sum_{k'=1}^{l-1} D'_{w_{k'}}/S_{w_{k'}})$$

(See Lemma 7 in Appendix D)

5: **else if** $flag_j = 1$, for one $j \in \{1, \ldots N_i\}$ **then** // Two femto split users case occurred in $\mathcal{F}_j(\alpha, \epsilon_i)$

6: Find $l \leq |W|$ such that $S_{w_l}/R_{w_l} > \alpha > S_{w_{l+1}}/R_{w_{l+1}}$

$$y_{w_k} = \begin{cases} D'_{w_k} & \text{for } 1 \leq k \leq l \\ 0 & \text{for } l+1 \leq k \leq |W| \end{cases}$$

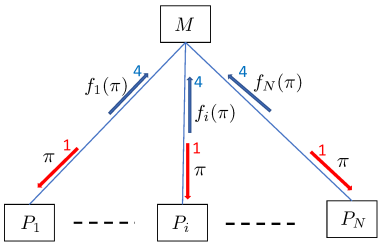
$$y_a = S_a(\pi - \epsilon_i - \sum\nolimits_{k=1}^{l} D'_{w_k}/S_{w_k})$$

7: $y_b = 0$, $z_a = D_a - y_a$, $z_b = T_b(\delta - z_a/T_a)$ (See Lemma 8 in Appendix D)

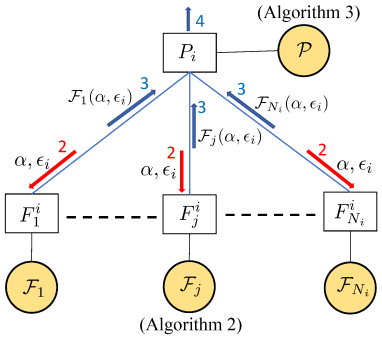
8: **end if**

9: **return** $\boldsymbol{y}, z_a, z_b$

A. Scalable and Distributed Implementation



(a) Macro level distributed scheme for computing clearing time. Here, $f_i(\pi)$ is computed as (6) using the results of $\Theta(\alpha, \epsilon_i)$ in Fig. 4(b)



(b) Pico level scheme - Distributed implementation of Algorithm 1 to get $\Theta(\alpha, \epsilon_i)$

Fig. 4: Distributed computation schemes. The yellow circles represent the local allocation functions at the BSs, and the arrows represent the messages. The numbering on the arrows is the order in which the message exchanges occur.

Fig. 4(b) shows a distributed implementation of Algorithm 1 to evaluate $\Theta(\alpha, \epsilon_i)$. The scheme can be implemented as follows. The pico P_i broadcasts a message containing the values (α, ϵ_i) to the femtos $\{F^i_j\}_{j=1}^{N_i}$. Then, each femto F^i_j runs the function $\mathcal{F}_j$ locally with the inputs (α, ϵ_i). The femto allocations $\{\mathcal{F}_j(\alpha, \epsilon_i)\}_{j=1}^{N_i}$ are computed in parallel at the corresponding femtos. Following the computation, each femto F^i_j sends the evaluation $\mathcal{F}_j(\alpha, \epsilon_i)$ to the pico P_i, which then computes $\mathcal{P}([\mathcal{F}_j(\alpha, \epsilon_i)]_{j=1}^{N_i})$. This completes the pico and femto allocations $\boldsymbol{z}, \boldsymbol{y}$. Macro allocation $\boldsymbol{x}$ is determined by line 3 of Algorithm 1.

This implementation is scalable in the number of femtos, N_i, due to the local nature of functions $\mathcal{F}_j$. The only increase is in the number of messages (equal to N_i) sent from the femtos F^i_j to the pico P_i. A similar statement about scalability also holds true for macro-level process shown in Fig. 4(a). The worst case computational complexity of the function $\mathcal{F}_j$ is $O((|U^i_j|+1)\log|U^i_j|)$, and for function $\mathcal{P}$, it is $O((|U_i|+1)\log|U_i|)$.

The only thing left is the search procedure to find the optimal values (α^*, ϵ^*_i). First, we introduce the notation necessary for discussion. Let $\theta(\alpha, \epsilon_i)$ be the value of the objective

function $\sum_{u\in U_i} x_u/R_u$ under the solution given by $\Theta(\alpha,\epsilon_i)$. If the solution is infeasible, we take $\theta(\alpha,\epsilon_i)$ to be ∞. Now, $(\alpha^*,\epsilon_i^*) := \arg\min_{\alpha\in A,\epsilon_i\in[0,\pi]} \theta(\alpha,\epsilon_i)$, and the optimal value of LP (5), $f_i(\pi)$ is given by

$$f_i(\pi) = \theta(\alpha^*,\epsilon_i^*) \tag{6}$$

The search algorithms and their convergence results are presented in the following section.

IV. Numerical Results

To illustrate the results, we consider a three tier HetNet with a macro BS, 6 pico BSs and 4 femto BSs per pico site. The simulation parameters are given in the following tables. The BS parameters are in the order: macro, pico, femto.

BS parameters	Values
Transmit power	46, 30, 22 (in dBm)
Antenna gain	14, 5, 3 (in dBi)
Path-loss exponent n	3.76, 3.76, 3
Coverage radius	500, 150, 50 (in m)
Log-normal shadowing standard deviation	10, 6, 6 (in dB)

Parameter	Value
Transmission bandwidth	10 MHz
File size D_u	2.7 Mb
UE noise figure	10 dB
Noise power	-106 dBm
Minimum inter-BS distance	300 m (for pico tier) 90 m (for femto tier)

The macro BS is placed at the origin, the other BS locations are randomly realized in the macro coverage region such that inter-BS distances are greater than the specified values. We consider circular cells with the specified radii; a UE receives signal from a BS if within the coverage radius. UE placement is done in two stages, 5 UEs are uniformly scattered in each femto cell in the first stage, and 20 UEs are uniformly scattered in each pico cell in the second stage. The path-loss (in dB) formula is $128 + 10n\log_{10}(d/\text{km})$, where d is the BS-UE distance.

A. Search for α^*, ϵ_i^*

In this section, we focus on the search to find ϵ_i^* and α^*. We start by fixing $\pi = 0.4$ sec in LP (1-4), and solve LP (5) by finding ϵ_i^* and α^*. Recall from (6) that $f_i(\pi) = \theta(\alpha^*,\epsilon_i^*)$.[4]

For a given ϵ_i, define $\alpha(\epsilon_i) := \arg\min_{\alpha\in A}\theta(\alpha,\epsilon_i)$ as the α that minimizes the objective function. We consider a layered search over α,ϵ_i. In section IV-A1, the inner search to find $\alpha(\epsilon_i)$ (shown in Fig. 6(a)). In section IV-A2, the outer search for ϵ_i^* (shown in Fig. 6(b)). Note that $\alpha^* = \alpha(\epsilon_i^*)$, hence both α^* and ϵ_i^* are derived here.

1) Inner search for $\alpha(\epsilon_i)$: We find the $\alpha(\epsilon_i)$ for the given ϵ_i using inner search in Fig. 6(a). Recall that $flag_j = 1$ is used to denote two split users case in Algorithm 2. Define $flag := \max_{j=1}^{N} flag_j$. One of the following conditions will hold when the input $\alpha = \alpha(\epsilon_i)$

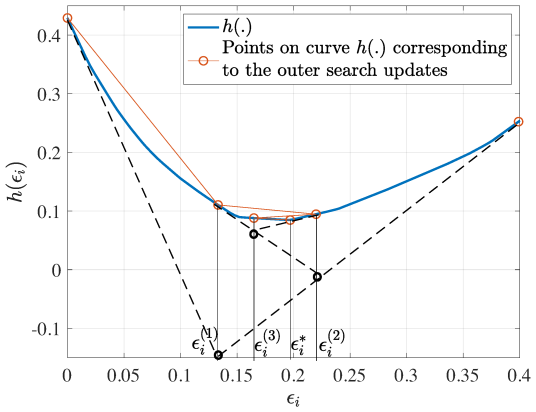
[4]There is macro-level search over π to minimize $\pi + \sum_{i=1}^{N} f_i(\pi)$ is presented in section IV-C. The value $\pi = 0.4 < \pi^*$ is chosen such that the constraints are tight, i.e., $f_i(\pi) > 0, \forall i$. The search is more straightforward when there is slackness.

i) If $flag = 0$, then $\alpha = S_{w_l}/R_{w_l}$, where w_l is the split user in Algorithm 3. (See Lemma 7 in Appendix D)

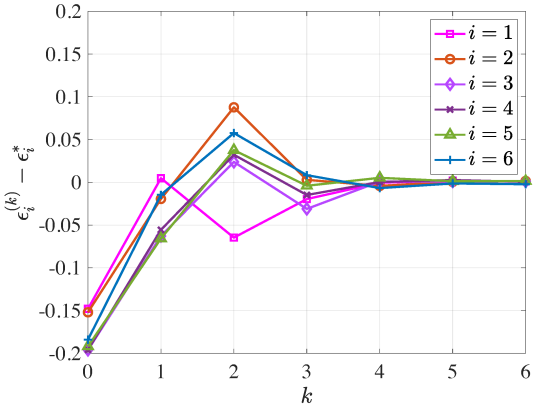
ii) If $flag = 1$, then $\alpha = S_aT_b/T_aR_b$, $0 \le y_a \le D_a$ and $0 \le z_b \le D_b$, where a, b are the two femto split users in Algorithms 2 & 3. (See Lemma 6 in Appendix C)

When conditions i) and ii) do not hold, either $\alpha > \alpha(\epsilon_i)$ (over-biased) or $\alpha < \alpha(\epsilon_i)$ (under-biased). Fig. 6(a) provides the criteria to check this relation between α and $\alpha(\epsilon_i)$, depending on the value of $flag$. Using this property, Fig. 6(a) performs a binary search for $\alpha(\epsilon_i)$ over A. In each iteration, $|A_{tmp}| \le |A'|/2$, since α is the median of set A'. Therefore, the set of possible α's is halved in size during the update $A' := A_{tmp}$. Hence, the convergence time (in steps) is at most $\log_2|A|$.

The average convergence times of inner search for the 6 picos are $[6.43, 5.12, 6.25, 4.25, 6.12, 6.12]$ steps respectively, where $|A|$ for the picos are $[82, 82, 80, 79, 76, 76]$ respectively. Here, the averages are calculated over the input ϵ_i's given by the outer search updates in Fig. 5(b).



(a) Outer search updates for $i = 1$

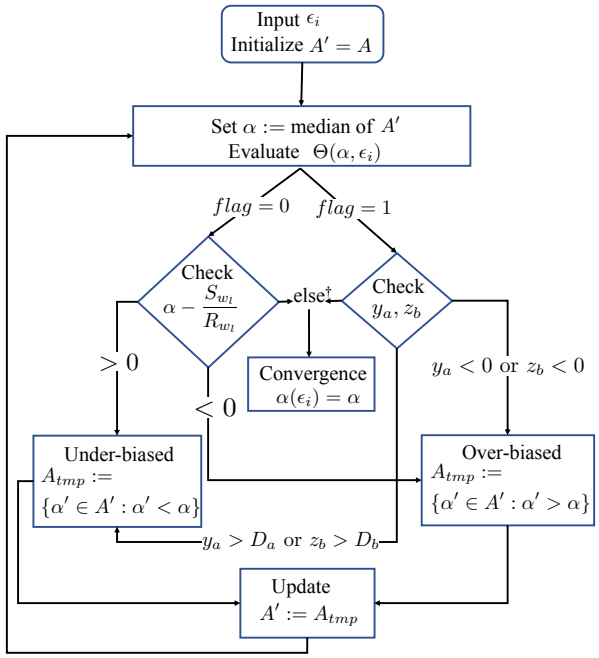


(b) Convergence of outer search for all the picos

Fig. 5: Outer search algorithm. Here k is the number of iterations, and $\epsilon_i^{(k)}$ is the value of ϵ_i in kth iteration.

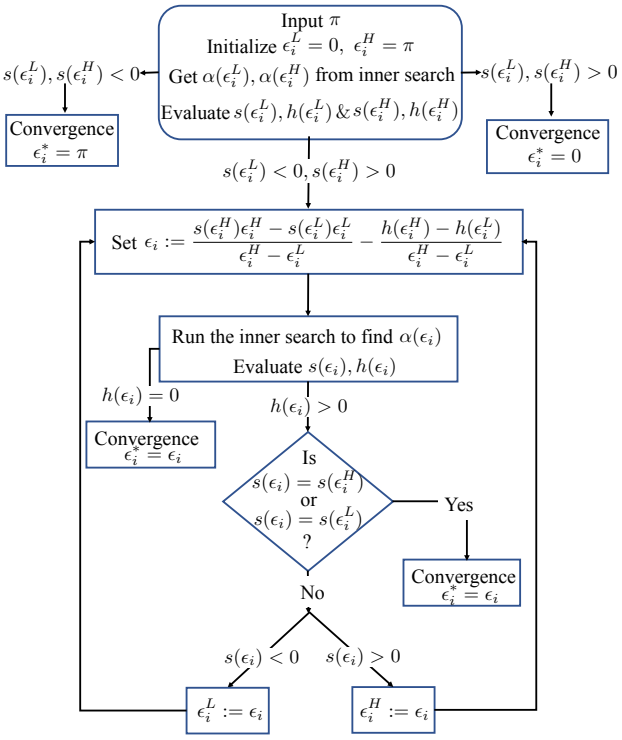
2) Outer search for ϵ_i^*: Define $h(\epsilon_i) := \theta(\alpha(\epsilon_i),\epsilon_i)$. Note that $h(\epsilon_i)$ is the value of LP (5) for a fixed given ϵ_i. We use the convexity of $h(.)$ to find $\epsilon_i^* := \arg\min_{\epsilon_i} h(\epsilon_i)$, using the outer search algorithm in Fig. 6(b).

Under the shadow-price interpretation of dual-variables, $s(\epsilon_i) := \partial h(\epsilon_i)/\partial\epsilon_i = \alpha' - \sum_{j=1}^{N_i}\beta_j'$, where $\alpha', \{\beta_j'\}_{j=1}^{N_i}$



† The else condition here is satisfied via the left arrow when $\alpha = S_{w_l}/R_{w_l}$, and via the right arrow when $0 \le y_a \le D_a$ & $0 \le z_b \le D_b$

(a) Inner search for $\alpha(\epsilon_i)$.



(b) Outer search for ϵ_i^*

Fig. 6: Search algorithms to find α^*, ϵ_i^*

are the dual-variables corresponding to the pico-time and femto-time constraints in LP (5). The rate bias multipliers $\alpha(\epsilon_i), \{\beta_j\}_{j=1}^{N_i}$ are equal to the corresponding dual-variables, provided the corresponding constraint is not slack. When a constraint is slack, the corresponding dual-variable is zero. (Refer to Appendix E for more details). Note that $\alpha(\epsilon_i)$ and $\Theta(\alpha(\epsilon_i), \epsilon_i)$ are evaluated by the inner search algorithm (in Fig. 6(a)). The gradient $s(\epsilon_i)$ can now be calculated since 1) $\alpha(\epsilon_i)$ is known, and 2) the allocation $[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}]$ (which determines slackness of constraints) and rate-multipliers $\boldsymbol{\beta}$ are given by $\Theta(\alpha(\epsilon_i), \epsilon_i)$.

$h(.)$ is a piecewise linear function (blue curve in Fig. 5(a)). Fig. 6(b) provides a linear interpolation based search algorithm to find ϵ_i^* in a finite number of steps. We take a point ϵ_i^L with a negative gradient and a point ϵ_i^H with a positive gradient and solve for a new ϵ_i as the ϵ_i-coordinate of the intersection of tangents (of curve $h(.)$) at points $(\epsilon_i^L, h(\epsilon_i^L))$ and $(\epsilon_i^H, h(\epsilon_i^H))$. e.g., In Fig 5(a), $\epsilon_i^L = 0, \epsilon_i^H = 0.4$ during iteration 1, and $\epsilon_i^{(1)}$ is the new ϵ_i. Now, either $\epsilon_i = \epsilon_i^*$ or the point $(\epsilon_i, h(\epsilon_i))$ lies on a new line segment of the curve $h(.)$ (See Fig. 5(a)). Due to convexity of $h(.)$, ϵ_i is closer to the ϵ_i^* than at least one of $\epsilon_i^L, \epsilon_i^H$. Finally, either ϵ_i^L or ϵ_i^H is updated based on the slope $s(\epsilon_i)$. The convergence occurs in finite number of steps because the curve $h(.)$ is composed of a finite number of line segments.

The convergence results can be seen in Fig 5. In Fig 5(a), ϵ_i^L is updated in iterations 1 and 3 (since the slope $s(\epsilon_i)$ is negative), and ϵ_i^H is updated in iteration 2. Fig. 5(b) shows the convergence times (in number of iterations or steps) for all the 6 picos.

B. Alternate approximate methods and convergence times

The search algorithms given in Fig. 6 in section IV-A derive the exact solution (α^*, ϵ_i^*) in finite number of steps. The simulation results indicate convergence with in a small number of steps. However, in practical implementation, issues like delay may impose additional constraints on search time. In this case, the search can be truncated and last calculated feasible solution can be used, which lies within $\epsilon_i^H - \epsilon_i^L$ distance of the optimal value ϵ_i^*.

Alternatively, we now present an approximate scheme with bounded convergence time (in steps). Here, the parameters α, ϵ_i are allowed to take values from a predefined finite set, e.g., quantized levels for parameters. Let S_α, S_{ϵ_i} denote the sets of values that α and ϵ_i can take respectively. We present the modified search algorithms as follows.

For the inner search, the algorithm in Fig. 6(a) can be applied with initialization $A' = S_\alpha$, and stopped when $|A'| = 1$. The convergence time is $\log_2 |S_\alpha|$. For the outer search, a binary search version of the algorithm in Fig. 6(b) can be applied, where the new $\epsilon_i \in S_{\epsilon_i}$ will be chosen as the median value between ϵ_i^L and ϵ_i^H (instead of the intersection of the tangents). Convergence occurs in $\log_2 |S_{\epsilon_i}|$ steps (when $\epsilon_i^L = \epsilon_i^H$). The total convergence time is $\le \log_2 |S_\alpha| \log_2 |S_{\epsilon_i}|$.

C. Performance Results of the Minimum Time Clearing Scheme

Recall that there is also a process at macro level (shown in Fig. 4(a)) to solve LP (1-4), i.e., to derive $\pi^* := \min_\pi \pi +$

$\sum_{i=1}^{N} f_i(\pi)$. A similar search method to Fig 6(b) or traditional methods such as golden section search, line search can be used to find π^*. Since, our main focus is on LP (5), we have only presented the convergence results for finding $f_i(\pi)$. Now, we present the clearing time $\pi + \sum_{i=1}^{N} f_i(\pi)$ as a function of π (red curve A in Fig. 7(a)), and compare with alternative schemes.

For comparison, we consider schemes A-D given in the following table. Scheme A is the minimum time clearing scheme of the paper, which uses full resource partitioning (FRP) between the tiers. Scheme D has no resource partitioning (No RP) between the tiers, and all the BSs are allowed to transmit simultaneously. For the other schemes B&C, we consider Almost Blanking Subframes (ABS) scheme of 3GPP. Under ABS, resource partitioning at macro tier is performed; the macro is silent during the small cell (or ABS) time. The picos and femtos are taken to use the entirety of small cell time for transmission. Scheme C uses SINR biased user association, which is equivalent to the Cell Range Expansion (CRE) scheme of 3GPP. The other schemes A,B&D use rate biased user association (as explained in the paper).

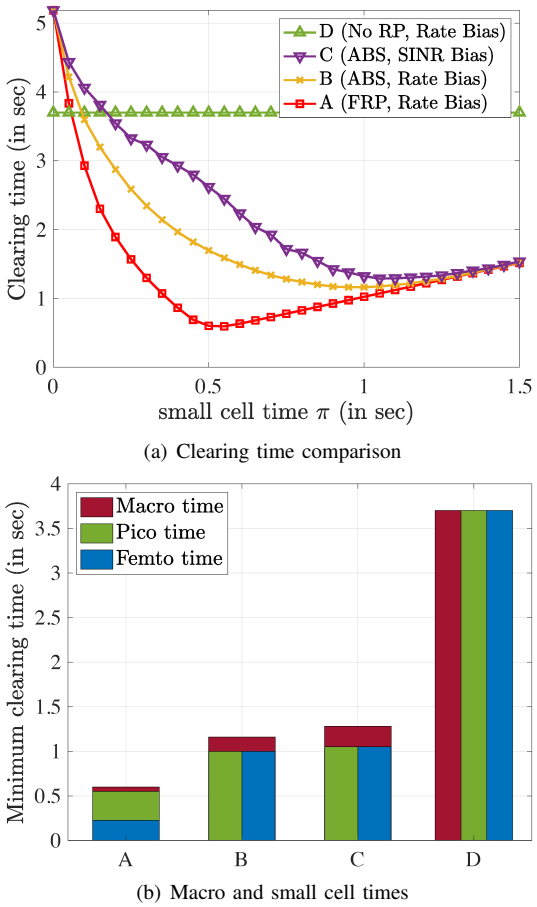
Note that the rates under No RP and ABS will be lower (than that of FRP) due to the cross-tier interference resulting from the simultaneous transmissions of different tiers.

Scheme	Resource partitioning	UE association rule
A	FRP	Rate Bias
B; C	ABS	Rate Bias; SINR Bias resp.
D	No RP	Rate Bias

We measure performance in terms of the time required to clear the files of a given set of UEs. Note that smaller clearing time means higher capacity, since more files are transmitted per second. The schemes A-C are adaptive with respect to the small cell time π, and hence the clearing time is minimized over all possible choices of bias values for each π. Scheme D is fixed. It is optimized over all possible bias values and has a fixed small cell time π (given by optimal biasing). Therefore, the clearing times presented are the best possible for respective schemes. Note that the clearing time of C provides a lower-bound to the CRE and ABS schemes of 3GPP. D is a lower-bound to the rate-biased schemes in [6], [16], [17], [20].

The results are presented in Fig. 7. It is clear that the minimum clearing scheme A performs better than the other schemes by definition. However, the difference is significant in the considered scenario, as can be observed from Fig. 7(a) and Fig. 7(b). It can also be observed that FRP (scheme A) provides significant gain over ABS (scheme B) for rate biased association, and the difference is even more significant between ABS (scheme B) and No RP (scheme D). This result highlights the importance of resource partitioning in HetNets.

Fig. 7(b) shows the distribution of macro, pico and femto times across the considered schemes at their respective optima. Here, pico time (and femto time) is the time available to the picos (and the femto resp.). Under ABS (B&C), the small-cell time π is available to all the picos and the femtos. For schemes B&C, we illustrate this with two parallel bars (green & blue). Under FRP, recall that the time available to a F_j^i is ϵ_i, i.e., it depends on i. For scheme A, the stacked green and blue



(a) Clearing time comparison

(b) Macro and small cell times

Fig. 7: Comparison of various user association and resource partitioning schemes.

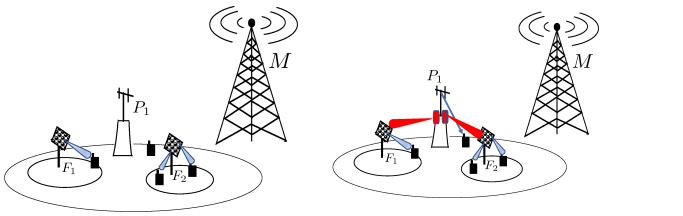
bars are the average pico time ($\pi - \sum_i \epsilon_i/N$) and femto time ($\sum_i \epsilon_i/N$) respectively. For D, the macro, pico and femto BSs are all transmitting at the same time, which is illustrated with three parallel bars (brown, green & blue). Scheme A has the smallest macro-cell load, followed by B & C. Lack of resource partitioning in D has resulted in a high macro-cell load.

V. Application to Future Networks

Thus far, the framework developed in the paper has been used to obtain the minimum time clearing scheme A (FRP, Rate Bias) in section IV-C. However, the framework is more general and can be easily adapted to implement other three tier joint optimization schemes. For example, the framework can be used to optimize three tier user association under an ABS setup as follows. Consider the three tier HeNet (described in section II), but now operating under a macro only ABS scheme (as in schemes B&C in section IV-C). Note that this cannot be addressed as a two-tier problem, since the picos and femtos must be distinguished for cell association. Recall that under ABS, the picos and femtos transmit during the entirety of small cell time π. UE SINRs (and rates) from femto F_j^i will now include the cross-tier interference from pico P_i and vice versa. With this setup, the minimum clearing time LP can be formulated as LP (1-4) with $\pi - \epsilon_i$ in (2) and ϵ_i in (3), replaced by π. The solution yields the optimal point of

scheme B (ABS, Rate Bias), i.e., the minimum of the yellow curve B in Fig. 7(a).

In this section, we provide a detailed application of the framework to an interesting future network, which will result in a different LP formulation (i.e., not LP (1-4)). We consider a three tier HetNet with mmWave femtos and mmWave wireless backhaul. We will show that all the main results can be extended to the mmWave HetNet, and describe how the algorithms can be implemented with slight modifications. For the mmWave cells, we consider single stream MIMO beamforming. The results can be extended to networks with advanced techniques such as Space Division Multiplexing (SDMA), but are beyond the scope of this paper.



(a) mmWave femto cell (b) mmWave femto cell with mmWave backhaul

Fig. 8: mmWave Three tier HetNets

Consider a 3 tier HetNet with the femto BSs using mmWave frequencies. Hence, the femto transmissions do not experience interference from the macro or the pico BSs, and require no radio resources from these cells. Time only needs to be partitioned between the macro and pico tier to avoid cross-tier interference. We consider a setup where the allocation is performed periodically at the beginning of each frame. The frame length is Δ seconds. Let $\pi \leq \Delta$ denote the time allocated to the small cells, which will now be used exclusively by the pico BSs. Therefore, the pico time constraint for P_i will now be $\sum_{u \in U_i} y_u/S_u \leq \pi$.

1) mmWave link rates and mmWave backhaul: We consider a setup where the femtos do not have a wired backhaul link. Pico BS P_i are equipped with special hardware to provide backhaul over a dedicated mmWave link to each F_j^i. The femto BSs employ beam-forming for serving UEs and also for backhaul.

Consider a UE $u \in U_j^i$ in the range of femto F_j^i. Let B_m denote the mmWave bandwidth available for serving UEs. Let $P_{F_j^i}$ denote the transmit power of the femto BS, and $g_{F_j^i,u}$ denote the gain of the link between F_j^i and the UE u, including the beam-forming directivity gain of the BS and the UE, path loss and shadowing loss. σ^2 is the noise power. The rate of the link between femto F_j^i and UE u is $T_u = B_m \log_2(1 + P_{F_j^i} g_{F_j^i,u}/(\sum_{k \neq j} P_{F_k^i} g_{F_k^i,u} + \sigma^2))$. Due to the directional nature of mmWave links, we take $\sum_{k \neq j} P_{F_k^i} g_{F_k^i,u}$ to be zero for the numerical simulations in this section. For the blocked UEs, T_u is zero - these UEs have to associate with a pico or macro BS. We assume that the rate T_u remains constant for the duration of the frame.

Let S_j^i denote the rate of the backhaul link between the pico P_i and femto F_j^i, calculated similarly as T_u. The backhaul link has to carry all the traffic into the femto F_j^i, i.e., $\sum_{u \in U_j^i} z_u$. Due to the half-duplex constraint, the frame has to be partitioned between the backhaul and UE transmissions of a femto. Therefore, the femto time constraint for F_j^i will now be $\sum_{u \in U_j^i} z_u/S_j^i + \sum_{u \in U_j^i} z_u/T_u \leq \Delta$. This is equivalent to $\sum_{u \in U_j^i} z_u/T'_u \leq \Delta$, where

$$T'_u = T_u/(1 + T_u/S_j^i) \tag{7}$$

2) Rate requirements: Let a_u (in bits/s) represent the rate requirement (or target) of a UE u. We assume that the rate requirements are set by a scheduler (based on some fairness criterion or a QOS requirement). The number of bits needed by u in the frame to meet the rate requirement is $a_u\Delta$. Now, the objective of minimizing the clearing time of the microwave part of HetNet is formulated as LP (8).

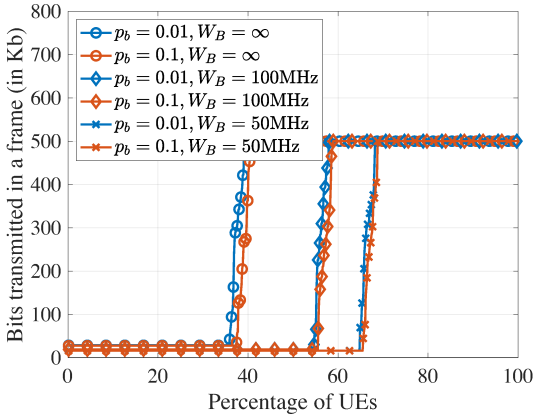
$$\begin{aligned} \min_{x_u, y_u, z_u, \pi \geq 0} \quad & \pi + \sum_{u \in \bigcup_{i=1}^{N} U_i} x_u/R_u \\ \text{s.t.} \quad & \sum_{u \in U_i} y_u/S_u \leq \pi, \forall i \in \{1, \dots N\} \\ & \sum_{u \in U_j^i} z_u/T'_u \leq \Delta, \forall j \in \{1, \dots N_i\}, i \in \{1, \dots N\} \\ & x_u + y_u + z_u = a_u\Delta, \forall u \in U \end{aligned} \tag{8}$$

Let $\bar{\Delta}$ be the value of LP (8). Note that when $\bar{\Delta} > \Delta$, the rate requirements cannot be met. We can take the solution $(\boldsymbol{x}^*, \boldsymbol{y}^*, \pi^*)$ and scale by $\Delta/\bar{\Delta}$ to get a feasible allocation for the current frame. Similarly when $\bar{\Delta} < \Delta$, scaling by $\Delta/\bar{\Delta}$ produces a maximal solution, so that no time is wasted in the frame.

3) Solution and Algorithms: Note that LP (8) is simpler to solve than LP (1-4) since it has one less variable, ϵ_i (we have a constant Δ instead). As before, fixing π, the problem can be decomposed into N independent LPs. The solution of each LP can be found by a 1D search over parameter α at P_i (i.e., inner search algorithm in section IV-A).

To apply the allocation algorithms given in the paper, firstly, it can be seen that T_u should be replaced with T'_u. Algorithm 2 and Algorithm 3 can now be applied by modifying the inputs. $\mathcal{F}_j(\alpha^*, \Delta)$ can be implemented as Algorithm 2 to solve for $\boldsymbol{z}^*$ and β_j^* at femto F_j^i. $\mathcal{P}(\alpha^*, 0, \boldsymbol{z}^*, a, b, \delta)$ can be implemented as Algorithm 3 to solve for $\boldsymbol{y}^*$.

4) Numerical Example: Due to the high rates of mmWave BSs, the value of S_j^i has a significant impact on T'_u (unlike the cases where $S_j^i \gg T_u$). To illustrate this effect, we consider the same BS setup as in section IV, with mmWave femtos. 30 UEs are uniformly placed within 50 m of each femto BS. The mmWave simulation parameters are given in Fig. 9(b). Here, the rate requirement a_u is the same (50 Mbps) for all the UEs. The blockage state (blocked or unblocked) of a UE is fixed according to the blockage probability p_b, prior to optimization. The numerical results can be seen in Fig. 9(a). The UEs receiving 50 Mbps (or 500 Kb/frame) are the mmWave UEs. Note that there is not enough bandwidth to support 50 Mbps rate for the microwave UEs, and hence they get scaled down rates.



(a) Cumulative distribution of throughput

Parameter	Value
Femto BS and backhaul Tx power	22 dBm
Femto BS directivity gain	20 dB
UE directivity gain	10 dB
UE throughput requirement a_u	50 Mbps
mmWave pathloss	3GPP UMi Model
mmWave femto bandwidth	50 MHz
mmWave backhaul bandwidth	W_B
Link blockage probability	p_b
Frame size Δ	10 ms

(b) mmWave simulation parameters

Fig. 9: Effect of backhaul bandwidth and blocking on mmWave HetNet capacity

From Fig. 9(a), it can be observed that changing blockage probability does not significantly change the solution of LP (8). This is because some UEs need to be offloaded to the microwave BSs anyway, and blocking just affects which ones are offloaded. In this case, adapting the bias-values based on the state of blocking only has a minor impact. The proposed scheme can therefore be implemented on a slower-time scale, with the blocked UEs changing association to microwave BSs using the given bias-values, and the mmWave femto swaps the blocked UE with an unblocked microwave UE (the one with the highest ρ_u^α).

Secondly, it can be seen from Fig. 9(a) that the backhaul bandwidth W_B has a significant impact on the traffic supported by the mmWave BSs. (7) shows that doubling S_j^i does not double T_u'. In Fig. 9(a), doubling W_B from 50 MHz to 100 MHz, only increased the number of femto UEs by $\approx 10\%$. We conclude that backhaul bandwidth needs to be accounted for in capacity planning, and there are diminishing returns from increasing it.

Appendix A: Proofs of Theorems 1,2 & 3

We develop the necessary theory in Appendices B-E. The results will be used in the proofs here. The layout of other Appendices is as follows. In Appendix B, we introduce the Lagrangian and the dual-variables corresponding to LP (5) and derive structural results of the optimal solution using the KKT conditions. In Appendix C, we present the femto allocation results using the KKT conditions, i.e., Lemmas supporting Algorithm 2. In Appendix D, we present the pico allocation results, i.e., Lemmas supporting Algortithm 3. The results of Appendices C-D were derived under the assumption that the dual variable $\alpha > 0$. In Appendix E, we deal with the case $\alpha = 0$.

Proof of Theorem 1. Suppose the dual-variables $\alpha, \beta_j > 0$, $\forall j \in \{1, \dots N_i\}$ for $\epsilon_i = \epsilon_i^*$. The proof of (1-3) of Theorem 1 follows from (9-12) of Appendix B. The results of Fig. 3 follow from Lemma 1 and Lemma 2 in Appendix B.

Suppose one or more of the dual-variables α, β_j are zero for $\epsilon_i = \epsilon_i^*$. It follows from Lemmas 9 & 10 in Appendix E, that there exist α_m and β_j^m corresponding to the zero dual-variables such that (1-3) of Theorem 1 hold. Now, the results of Fig. 3 follow from Lemma 1 and Lemma 2 with α (and β_j) replaced with α_m (and β_j^m resp.). □

Proof of Theorem 2. Suppose the dual variable $\alpha > 0$ for $\epsilon_i = \epsilon_i^*$. If Pico allocation case 1 (in Appendix D) holds, then $\alpha = S_{w_l}/R_{w_l}$ is an optimal dual-variable from Lemma 7. If Pico allocation case 2 (in Appendix D) holds, then $\alpha = S_a T_b / T_a R_b$ from Lemma 8.

Suppose the dual variable $\alpha = 0$ for $\epsilon_i = \epsilon_i^*$. It follows from Lemma 9 in Appendix E that $\exists \alpha_m \in A$ such that $\theta(\alpha_m, \epsilon_i) = 0$. Here, $\alpha^* = \alpha_m$.

Similarly, suppose the dual-variable $\beta_j > 0$ for $\epsilon_i = \epsilon_i^*$. Then $\beta_j^* = \rho_u^\alpha$ from Appendix C. Otherwise if dual-variable $\beta_j = 0$, $\beta_j^m = \min_{u \in U_j^i} \rho_u^\alpha$ is the rate-bias multiplier β_j^* from Lemma 10 in Appendix E. □

Proof of Theorem 3. Suppose the dual variable $\alpha > 0$ for $\epsilon_i = \epsilon_i^*$. It follows from Lemmas 4-6 that Algorithm 2 determines $\boldsymbol{z}^*$ with α, ϵ_i^* as input (See Appendix C). It follows from Lemmas 7-8 that Algorithm 3 determines $\boldsymbol{y}^*$. (See Appendix D)

If the dual variable $\alpha = 0$, the proof follows from Lemma 9 in Appendix E. □

Appendix B: KKT conditions and Lagrangian minimization

For the given π, we start by fixing a $\epsilon_i \in [0, \pi]$. We consider LP (5) for the given π, ϵ_i[5].

Consider the Lagrangian L of LP (5), given as

$$L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{u \in U_i} x_u / R_u + \alpha \Big(\sum_{u \in U_i} y_u / S_u + \epsilon_i - \pi \Big)$$

$$+ \sum_{j=1}^{N_i} \beta_j \Big(\sum_{u \in U_j^i} z_u / T_u - \epsilon_i \Big) - \sum_{u \in U_i} \gamma_u (x_u + y_u + z_u - D_u)$$

where α, β_j and γ_u are the dual variables corresponding to the constraints of LP (5) (see page 3).

For a fixed (π, ϵ_i), LP (5) is equivalent to the Lagrangian minimization problem $\min_{x_u, y_u, z_u \geq 0} L$ with the optimal dual-variables. The KKT conditions provide sufficient conditions for optimality of the primal and dual variables.

[5]Let $g(\pi, \epsilon_i)$ denote the solution of LP (5) for the given pair (π, ϵ_i). Note that $f_i(\pi) = \min_{\epsilon_i \in [0,\pi]} g(\pi, \epsilon_i)$, and $\epsilon_i^* = \arg\min_{\epsilon_i \in [0,\pi]} g(\pi, \epsilon_i)$.

A. Stationarity conditions

From the first order stationarity conditions of the KKT conditions, we must have

$$\partial L/\partial x_u = 1/R_u - \gamma_u \geq 0 \tag{9}$$

and $\gamma_u = 1/R_u$ if $x_u > 0$. i.e., minimum occurs either at a stationary point or at a point on the boundary. Similarly, we have

$$\partial L/\partial y_u = \alpha/S_u - \gamma_u \geq 0 \text{ and } \gamma_u = \alpha/S_u \text{ if } y_u > 0 \tag{10}$$

$$\partial L/\partial z_u = \beta_j/T_u - \gamma_u \geq 0 \text{ and } \gamma_u = \beta_j/T_u \text{ if } z_u > 0 \tag{11}$$

Using (9-11), $\gamma_u \leq \min\{1/R_u, \alpha/S_u, \beta_j/T_u\}$. Since $x_u + y_u + z_u = D_u > 0$, at least one of $x_u, y_u, z_u > 0$ and hence, at least one of the equality conditions of (9-11) must hold. Therefore,

$$\gamma_u = \min\{1/R_u, \alpha/S_u, \beta_j/T_u\} \geq 0 \tag{12}$$

Going forward, we take $[x_u, y_u, z_u]_{u\in U_i}$ to be the solution of LP (5) for the given fixed π, ϵ_i, and $\alpha, \beta_j, \gamma_u$ to be the optimal dual variables, i.e., KKT conditions hold for these values.

Assumption 1. *For any $u, v \in U_i$ and $u \neq v$, we assume that 1) $T_u/S_u \neq T_v/S_v$, 2) $S_u/R_u \neq S_v/R_v$ and 3) $T_u/R_u \neq T_v/R_v$. Furthermore, for any $(u_1, v_1) \neq (u_2, v_2) \in U_i \times U_i$, we assume that $S_{u_1}T_{v_1}/R_{v_1}T_{u_1} \neq S_{u_1}T_{v_1}/R_{v_1}T_{u_1}$.*

Note that the rates R_u, S_u, T_u are arbitrary real values, and Assumption 1 holds with probability 1. For the sake of brevity, we ignore the highly special cases where Assumption 1 does not hold, e.g., a case where two different UEs have exactly the same rate-ratios mentioned in Assumption 1.

B. Lemmas on relationship between primal and dual variables

Recall that $\rho_u^\alpha := \min\{T_u/R_u, \alpha T_u/S_u\}$ from Theorem 1. Proofs of the following three lemmas are direct consequences of the stationarity conditions (9-12).

Lemma 1. *Suppose the dual variable $\alpha > 0$. Then 1) $z_u = D_u$, if $\rho_u^\alpha > \beta_j$ and 2) $z_u = 0$, if $\rho_u^\alpha < \beta_j$.*

Proof. Suppose $\rho_u^\alpha > \beta_j$. This implies $\beta_j/T_u < \min\{1/R_u, \alpha/S_u\}$. From (12), we have $\gamma_u = \beta_j/T_u$ and $\gamma_u < \min\{1/R_u, \alpha/S_u\}$. Now from (9-10), we have $x_u = 0, y_u = 0$, and hence $z_u = D_u$. Therefore, $\rho_u^\alpha > \beta_j$ implies $z_u = D_u$.

Now suppose $\rho_u^\alpha < \beta_j$. This implies $\beta_j/T_u > \min\{1/R_u, \alpha/S_u\}$. From (12), $\beta_j/T_u > \gamma_u$. Now from (11), we have $z_u = 0$. Therefore, $\rho_u^\alpha < \beta_j$ implies $z_u = 0$. □

Lemma 2. *Suppose $D'_u := D_u - z_u > 0$ for some $u \in U_i$. Then 1) $y_u = D'_u$, if $S_u/R_u > \alpha$ and 2) $y_u = 0$, if $S_u/R_u < \alpha$.*

Proof. Suppose $S_u/R_u > \alpha$. This implies $1/R_u > \alpha/S_u$, and hence $\gamma_u < 1/R_u$ from (12). We have $x_u = 0$ from (9). Therefore, $y_u + z_u = D_u$. This proves 1).

For 2), suppose $S_u/R_u < \alpha$. This implies $1/R_u < \alpha/S_u$, and hence $\gamma_u < \alpha/S_u$ from (12). Therefore, $y_u = 0$ from (10). □

Lemma 3. *Consider a user $u \in U_j^i$. 1) If $x_u, y_u > 0$, then $\alpha = S_u/R_u$. 2) If $y_u, z_u > 0$, then $\beta_j = \rho_u^\alpha = \alpha T_u/S_u$ and 3) If $z_u, x_u > 0$, then $\beta_j = \rho_u^\alpha = T_u/R_u$.*

Proof. Suppose $x_u, y_u > 0$. From (9-11), we have $\gamma_u = 1/R_u$, $\gamma_u = \alpha/S_u$. Therefore, $\alpha = S_u/R_u$. This proves 1).

Suppose $y_u, z_u > 0$. From (9-11), we have $\gamma_u = \alpha/S_u$, $\gamma_u = \beta_j/T_u$. Therefore, $\beta_j = \alpha T_u/S_u$

3) can be proved using similar arguments. □

APPENDIX C: FEMTO ALLOCATION

In this section, we present the femto allocation $[z_u]_{u\in U_j^i}$ for an arbitrary $j \in \{1, \dots N_i\}$. We will show that Algorithm 2 determines the femto allocation. This is done under the assumption that the dual-variable $\alpha > 0$. The other case $\alpha = 0$ is done in Appendix E.

Assume $\alpha > 0$. Recall that $\rho_u^\alpha := \min\{T_u/R_u, \alpha T_u/S_u\}$. Sort users u_k in U_j^i in descending order of $\rho_{u_k}^\alpha$ such that $\rho_{u_1}^\alpha \geq \rho_{u_2}^\alpha \geq \dots \geq \rho_{u_K}^\alpha$. Here $K = |U_j^i|$. Note that $\rho_{u_{k_1}}^\alpha = \rho_{u_{k_2}}^\alpha$ for at most one pair k_1, k_2 such that $1 \leq k_1 < k_2 \leq K$ (otherwise Assumption 1 is violated). Therefore, exactly one of the following three cases must hold

Case 1 (No split user case): Here (13) holds

$$\sum_{k=1}^{K} D_{u_k}/T_{u_k} \leq \epsilon_i \tag{13}$$

The femto allocation for this case is given in Lemma 4, which justifies step 3 of Algorithm 2.

Case 2 (Single split user case): $\exists l \leq K$ such that

$$1)\ \sum_{k=1}^{l-1} D_{u_k}/T_{u_k} \leq \epsilon_i < \sum_{k=1}^{l} D_{u_k}/T_{u_k} \tag{14}$$

$$2)\ \rho_{u_l}^\alpha \neq \rho_{u_k}^\alpha, \forall k \in \{1, 2, \dots K\} - \{l\} \tag{15}$$

The femto allocation for this case is given in Lemma 5, which justifies step 4 of Algorithm 2.

Case 3 (Two split users case): $\exists l \leq K-1$ such that

$$1)\ \sum_{k=1}^{l-1} D_{u_k}/T_{u_k} \leq \epsilon_i < \sum_{k=1}^{l+1} D_{u_k}/T_{u_k} \tag{16}$$

$$2)\ \rho_{u_l}^\alpha = \rho_{u_{l+1}}^\alpha \tag{17}$$

For the two split users, w.l.o.g assume $\rho_{u_l}^\alpha = \alpha T_{u_l}/S_{u_l}$ and $\rho_{u_{l+1}}^\alpha = T_{u_{l+1}}/R_{u_{l+1}}$. Define $a := u_l$, $b := u_{l+1}$ and $\delta := \epsilon_i - \sum_{k=1}^{l-1} D_{u_k}/T_{u_k}$. Here, a is the pico-femto split user and b is the macro-femto split user.

The femto allocation for this case is given in Lemma 6, which justifies step 5 of Algorithm 2. Note that z_a, z_b are not given by Lemma 6, and will be given in Lemma 8.

Lemma 4. *Suppose $\alpha > 0$ and (13) holds. Then $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq K$ and $\beta_j = 0$.*

Proof. We use proof by contradiction to show that $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq K$. Suppose not and assume $z_{u_p} < D_{u_p}$

for some $1 \leq p \leq K$. Since $z_{u_k} \leq D_{u_k}, \forall k \neq p$, we have $\sum_{k=1}^{K} z_{u_k}/T_{u_k} < \sum_{k=1}^{K} D_{u_k}/T_{u_k} \leq \epsilon_i$ from (13). Therefore $\beta_j = 0$ from complementary slackness. If $\beta_j = 0$, we have $z_{u_p} = D_{u_p}$ from Lemma 1, which is a contradiction to the assumption. Hence, $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq K$.

Now for determining β_j, if the inequality in (13) is strict, we have $\sum_{k=1}^{K} z_{u_k}/T_{u_k} < \epsilon_i$, and complementary slackness implies $\beta_j = 0$. Otherwise, if the equality in (13) holds, the KKT conditions hold for any $\beta_j \in [0, \rho^{\alpha}_{u_K}]$. □

Lemma 5. *Suppose $\alpha > 0$ and $\exists l \leq K$ such that (14),(15) hold. Then $\beta_j = \rho^{\alpha}_{u_l}$ and*

$$z_{u_k} = \begin{cases} D_{u_k} & \text{for } 1 \leq k \leq l-1 \\ T_{u_l}(\epsilon_i - \sum_{k'=1}^{l-1} D_{u_{k'}}/T_{u_{k'}}) & \text{for } k = l \\ 0 & \text{for } l+1 \leq k \leq K \end{cases}$$

Proof. 1) Firstly, we show that $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq l-1$ using proof by contradiction.

Suppose not and assume $z_{u_p} < D_{u_p}$ for some $1 \leq p \leq l-1$. Note that $\rho^{\alpha}_p \leq \beta_j$ from Lemma 1. This implies $\beta_j \geq \rho^{\alpha}_p > \rho^{\alpha}_l$ from (15). Therefore, $z_{u_k} = 0, \forall l \leq k \leq K$ from Lemma 1. This implies $\sum_{k=1}^{K} z_{u_k}/T_{u_k} < \sum_{k=1}^{l-1} D_{u_k}/T_{u_k} \leq \epsilon_i$ from (14). Therefore, $\beta_j = 0$ from complementary slackness. Observe that $\beta_j = 0$ implies $z_{u_p} = D_{u_p}$ from Lemma 1, which is a contradiction to the assumption $z_{u_p} < D_{u_p}$. Hence, $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq l-1$.

2) Now, we show that $z_{u_k} = 0, \forall l+1 \leq k \leq K$ using proof by contradiction.

Suppose not, and assume $z_{u_p} > 0$ for some $l+1 \leq p \leq K$. This implies $\beta_j \leq \rho^{\alpha}_p$. Therefore, $\rho^{\alpha}_{u_l} > \rho^{\alpha}_p \geq \beta_j$ from (15). Therefore, $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq l$ from Lemma 1. This implies $\sum_{k=1}^{K} z_{u_k}/T_{u_k} \geq \sum_{k=1}^{l} D_{u_k}/T_{u_k} > \epsilon_i$ from (14). This violates the primal constraint that $\sum_{k=1}^{K} z_{u_k}/T_{u_k} \leq \epsilon_i$, which is a contradiction. Hence, $z_{u_k} = 0, \forall l+1 \leq k \leq K$.

3) We now show that $\beta_j > 0$ and determine z_{u_l}.

Suppose not, and assume $\beta_j = 0$. We have $z_{u_l} = D_{u_l}$ from Lemma 1, which implies $\sum_{k=1}^{K} z_{u_k}/T_{u_k} > \epsilon_i$ from (14). This is a contradiction since it violates the primal constraint that $\sum_{k=1}^{K} z_{u_k}/T_{u_k} \leq \epsilon_i$. Hence, $\beta_j > 0$, which implies $\sum_{k=1}^{K} z_{u_k}/T_{u_k} = \epsilon_i$ from complementary slackness. Substituting the other values, $z_{u_l} = T_{u_l}(\epsilon_i - \sum_{k=1}^{l-1} D_{u_k}/T_{u_k})$

Note that when the left inequality of (14) is strict, $0 < z_{u_l} < D_{u_l}$, which implies $\beta_j = \rho^{\alpha}_{u_l}$ from Lemma 3. Otherwise, if the equality holds in the left inequality of (14), the KKT conditions hold for any $\beta_j \in [\rho^{\alpha}_{u_l}, \rho^{\alpha}_{u_{l-1}}]$. □

Lemma 6. *Suppose $\alpha > 0$ and $\exists l \leq K-1$ such that (16),(17) hold. Let $a := u_l$, $b := u_{l+1}$ and $\delta := \epsilon_i - \sum_{k=1}^{l-1} D_{u_k}/T_{u_k}$. W.l.o.g, let $\rho^{\alpha}_a = \alpha T_a/S_a$ and $\rho^{\alpha}_b = T_b/R_b$. Then*

$$z_{u_k} = \begin{cases} D_{u_k} & \text{for } 1 \leq k \leq l-1 \\ 0 & \text{for } l+2 \leq k \leq K \end{cases}$$

$z_a/T_a + z_b/T_b = \delta$, $\alpha = S_a T_b/T_a R_b$, and $\beta_j = \rho^{\alpha}_a$.

Proof. It can proved that $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq l-1$ and $z_{u_k} = 0, \forall l+2 \leq k \leq K$ using similar arguments as in the proof of Lemma 5.

Note that $\rho^{\alpha}_a = \rho^{\alpha}_b$ implies $\alpha T_a/S_a = T_b/R_b$. Hence $\alpha = S_a T_b/R_b T_a$.

It can proved that $\beta_j > 0$ using similar arguments as in the proof of Lemma 5. From complementary slackness, $\sum_{k=1}^{K} z_{u_k}/T_{u_k} = \epsilon_i$. Substituting other z values, we have $z_a/T_a + z_b/T_b = \delta$.

Note that $\delta < D_a/T_a + D_b/T_b$ from the right inequality of (16). This implies either $z_a < D_a$ or $z_b < D_b$. Therefore, $\beta_j \geq \rho^{\alpha}_a = \rho^{\alpha}_b$ from Lemma 1. Suppose the left inequality of (16) is strict, then $\delta > 0$. This implies at least one of $z_a, z_b > 0$, we have $\beta_j \leq \rho^{\alpha}_a = \rho^{\alpha}_b$ from Lemma 1. Therefore, $\beta_j = \rho^{\alpha}_a$ when the left inequality of (16) is strict. Otherwise, if the equality holds, the KKT conditions hold for any $\beta_j \in [\rho^{\alpha}_a, \rho^{\alpha}_{l-1}]$. □

Appendix D: Pico Allocation

We will present the pico allocation $[y_u]_{u \in U_i}$ under the assumption $\alpha > 0$. The other case $\alpha = 0$ is done in Appendix E. Let $D'_u := D_u - z_u$ denote the residual file after the femto allocation for $u \in U_i - \{a, b\}$. Recall that a, b are the split users from (16 -17) in Appendix C.

Let W denote the set of $u \in U_i - \{a, b\}$ such that $D'_u > 0$, i.e, positive residual file sizes after femto allocation. Sort the users $w_k \in W$ such that $S_{w_1}/R_{w_1} > \ldots > S_{w_{|W|}}/R_{w_{|W|}}$. We determine the pico allocation $[y_{w_k}]_{k=1}^{|W|}, y_a, y_b$ as the following two cases.

A. Pico allocation case 1

Suppose conditions (16-17) do not hold for any $j \in \{1, \ldots N_i\}$, i.e., Case 3 (Two split users case) in Appendix C does not hold for any j. Here, $\{a, b\} = \phi$. Since $\alpha > 0$, we have $\sum_{u \in U_i} y_u/S_u = \sum_{k=1}^{|W|} y_{w_k}/S_{w_k} = \pi - \epsilon_i$ from complementary slackness. Therefore, $\exists l \leq |W|$ such that

$$\sum_{k=1}^{l-1} D'_{w_k}/S_{w_k} < \pi - \epsilon_i \leq \sum_{k=1}^{l} D'_{w_k}/S_{w_k} \tag{18}$$

The following lemma provides the pico allocation for this case and justifies step 5 of Algorithm 3.

Lemma 7. *Suppose $\alpha > 0$ and $\{a, b\} = \phi$. Also, suppose that (18) holds for $l \leq |W|$. Then*

$$y_{w_k} := \begin{cases} D'_{w_k} & \text{for } 1 \leq k \leq l-1 \\ S_{w_l}(\pi - \epsilon_i - \sum_{k'=1}^{l-1} \frac{D'_{w_{k'}}}{S_{w_{k'}}}) & \text{for } k = l \\ 0 & \text{for } l+1 \leq k \leq |W| \end{cases}$$

Moreover, $\alpha = S_{w_l}/R_{w_l}$.

Proof. Firstly, we show that $y_{w_k} = D'_{w_k}, \forall 1 \leq k \leq l-1$, using proof by contradiction.

Suppose not, and assume $y_{w_p} < D'_{w_p}$ for some $1 \leq p \leq l-1$. Note that this implies $S_{w_p}/R_{w_p} \leq \alpha$ from Lemma 2. Therefore, $S_{w_k}/R_{w_k} < \alpha, \forall p+1 \leq k \leq |W|$. Therefore, $y_{w_k} = 0, \forall l \leq k \leq |W|$. This implies $\sum_{k=1}^{|W|} y_{w_k}/S_{w_k} \leq \sum_{k=1}^{l-1} D'_{w_k}/S_{w_k} < \pi - \epsilon_i$ from (18). Therefore, $\alpha = 0$ from complementary slackness, which implies $y_{w_p} = D'_{w_p}$ from Lemma 2. This is a contradiction to the assumption $y_{w_p} < D'_{w_p}$. Hence, $y_{w_k} = D'_{w_k}, \forall 1 \leq k \leq l-1$.

Now, we show that $y_{w_k} = 0, \forall l+1 \leq k \leq |W|$, using proof by contradiction.

Suppose not, and assume $y_{w_p} > 0$ for some $l+1 \leq p \leq |W|$. Note that this implies $S_{w_p}/R_{w_p} \geq \alpha$ from Lemma 2. Therefore, $S_{w_k}/R_{w_k} > \alpha, \forall 1 \leq k \leq p-1$. Since $l \leq p-1$, $y_{w_k} = D'_{w_k}, \forall 1 \leq k \leq l$ from Lemma 2, which implies $\sum_{k=1}^{|W|} y_{w_k}/S_{w_k} \geq \sum_{k=1}^{l} D'_{w_k}/S_{w_k} + y_{w_p}/S_{w_p} > \pi - \epsilon_i$ from (18). This violates the primal constraint that $\sum_{k=1}^{|W|} y_{w_k}/S_{w_k} \leq \pi - \epsilon_i$, which is a contradiction.

Now we determine α. If the right inequality in (18) is strict, then $0 < y_{w_l} < D'_{w_l}$ and $x_{w_l} > 0$. Hence, $\alpha = S_{w_l}/R_{w_l}$ from Lemma 3. Otherwise, if the equality holds in (18), the KKT conditions hold for any $\alpha \in [S_{w_{l+1}}/R_{w_{l+1}}, S_{w_l}/R_{w_l}]$. □

B. Pico allocation case 2

Suppose conditions (16-17) of Case 3 (Two split users) hold for some $j \in \{1, \dots N_i\}$ (See Appendix C). Lemma 8 provides the pico allocation for this case, and justifies step 7 of Algorithm 3.

Lemma 8. *Suppose $\alpha > 0$ and conditions (16-17) of Case 3 (Two split users) hold for some $j \in \{1, \dots N_i\}$ (See Appendix C). Then $\alpha = S_a T_b / T_a R_b$ and*

$$y_{w_k} = \begin{cases} D'_{w_k} & \text{for } 1 \leq k \leq l \\ 0 & \text{for } l+1 \leq k \leq |W| \end{cases}$$

$$y_a = S_a(\pi - \epsilon_i - \sum_{k=1}^{l} D'_{w_k}/S_{w_k})$$

$y_b = 0$, $z_a = D_a - y_a$ & $z_b = T_b(\delta - z_a/T_a)$. Further, (16-17) do not hold for any $j' \neq j$.

Proof. Note that $\alpha = S_a T_b / T_a R_b$ from Lemma 6. Due to Assumption 1, $S_{w_k}/R_{w_k} \neq \alpha, \forall 1 \leq k \leq |W|$. Therefore, $\exists l \leq |W|$ such that $S_{w_k}/R_{w_k} < \alpha, \forall 1 \leq k \leq l$ and $S_{w_k}/R_{w_k} > \alpha, \forall l+1 \leq k \leq |W|$. Therefore, $y_{w_k} = D'_{w_k}, \forall 1 \leq k \leq l$ and $y_{w_k} = 0, \forall l+1 \leq k \leq |W|$ from Lemma 2.

It remains to determine y_a, z_a, y_b, z_b. Recall from Lemma 6 that $\rho_b^\alpha = T_b/R_b < \alpha T_b/S_b$, which implies $S_b/R_b < \alpha$. Therefore, $y_b = 0$ from Lemma 2.

Since $\alpha > 0$, we have $\sum_{u \in U_i} y_u/S_u = \pi - \epsilon_i$ from complementary slackness. Substituting other y values, we get $y_a = S_a(\pi - \epsilon_i - \sum_{k=1}^{l} D'_{w_k}/S_{w_k})$.

For determining z_a, z_b, recall that $\rho_a^\alpha = \alpha T_a/S_a < T_a/R_a$ and hence $1/R_a > \gamma_a$ from (12). Therefore, $x_a = 0$ and $z_a = D_a - y_a$ from (9). Now, z_b can be determined from $z_a/T_a + z_b/T_b = \delta$.

Lastly, we prove that (16-17) do not hold for any $j' \neq j$. Suppose not and assume (16-17) holds for some $j' \in \{1, \dots N_i\} - \{j\}$. It follows from Lemma 6 that $\exists (a', b') \neq (a, b)$ such that $\alpha = S_a T_b / T_a R_b = S_{a'} T_{b'} / T_{a'} R_{b'}$, which violates Assumption 1. □

APPENDIX E: ZERO VALUED DUAL VARIABLES

In Appendices B-D, we have established that $\Theta(\alpha, \epsilon_i)$ determines the solution of LP (5) for any π, ϵ_i; provided the dual variable $\alpha > 0$. Here, α is also the pico rate bias multiplier. In this Appendix, we will show that there exist positive rate bias multipliers (such that (9-12) hold) when the corresponding dual-variables are zero.

Lemma 9. *If the dual variable $\alpha = 0$, $\exists$ a positive rate bias multiplier $\alpha_m \in A$ such that $\Theta(\alpha_m, \epsilon_i)$ gives an optimal allocation, and (9-12) hold when α is replaced with α_m.*

Proof. Since $\alpha = 0$; $\gamma_u = 0, \forall u \in U_i$ from (12). This implies $1/R_u > \gamma_u$ and $x_u = 0, \forall u \in U_i$ from (9). Therefore, the optimal value of the LP (5) is 0, and LP (19) must have a value $\leq \pi$ under the optimal solution. Note that LP (19) is a two-tier LP, which was considered in [25].

$$\begin{aligned} \min_{y_u, z_u \geq 0} & \sum_{u \in U_i} y_u/S_u \\ \text{s.t.} & \sum_{u \in U_j^i} z_u/T_u \leq \epsilon_i, \forall j \in \{1 \dots N_i\} \\ & y_u + z_u = D_u, \forall u \in U_i \end{aligned} \tag{19}$$

Define $\alpha_m := \min_{v \in U_i} S_v/R_v$ and note that $S_u/R_u \geq \alpha_m, \forall u \in U_i$. Therefore, $\alpha_m T_u/S_u \leq T_u/R_u$ which implies $\rho_u^{\alpha_m} = \alpha_m T_u/S_u, \forall u \in U_i$. Notice that when α_m is given as an input to $\Theta(\alpha_m, \epsilon_i)$, the UEs $u \in U_j^i$ will be sorted according to T_u/S_u in Algorithm 2 for each $j \in \{1, \dots N_i\}$. The femto allocation in this case coincides with the optimal two tier solution in [25]. Hence, $\Theta(\alpha_m, \epsilon_i)$ solves LP (19) and produces an allocation which satisfies $x_u = 0, \forall u \in U_i$. Since the value of the objective function $\sum_{u \in U_i} x_u/R_u = 0$, the solution is optimal. Hence, (9-12) hold under this optimal solution with the pico rate bias multiplier α_m in place of α. □

Lemma 10. *If the dual variable $\beta_j = 0$, $\exists$ a postive rate bias multiplier β_j^m such that (9-12) hold when β_j is replaced with β_j^m.*

Proof. Recall that when $\beta_j = 0$ in Lemma 4, $z_{u_k} = D_{u_k}, \forall 1 \leq k \leq K$. Consider $\beta_j^m := \rho_{u_K}^\alpha$. Since $\rho_{u_k}^\alpha \geq \beta_j^m, \forall 1 \leq k \leq K$, it follows $\beta_j^m/T_{u_k} < \min\{1/R_{u_k}, \alpha/S_{u_k}\}$. Hence, the rate-bias rules (9-12) hold when β_j is replaced with β_j^m. □

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152–1163, 2014.

[3] S. Chandrasekharan, K. Gomez, A. Al-Hourani, S. Kandeepan, T. Rasheed, L. Goratti, L. Reynaud, D. Grace, I. Bucaille, T. Wirth *et al.*, "Designing and implementing future aerial communication networks," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 26–34, 2016.

[4] M. M. Azari, F. Rosas, A. Chiumento, and S. Pollin, "Coexistence of terrestrial and aerial users in cellular networks," in *2017 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2017, pp. 1–6.

[5] A. Mohammed, A. Mehmood, F.-N. Pavlidou, and M. Mohorcic, "The role of high-altitude platforms (haps) in the global wireless connectivity," *Proceedings of the IEEE*, vol. 99, no. 11, pp. 1939–1953, 2011.

[6] 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, 2013.

[7] S. Borst, S. Hanly, and P. Whiting, "Throughput utility optimization in hetnets," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[8] C. Liu, M. Li, S. V. Hanly, and P. Whiting, "Joint downlink user association and interference management in two-tier hetnets with dynamic resource partitioning," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 2, pp. 1365–1378, 2016.

[9] X. Ge, X. Li, H. Jin, and J. Cheng, "Joint user association and user scheduling for load balancing in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3211–3225, 2018.

[10] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3388–3401, 2016.

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

[12] 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, 2013.

[13] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in hetnets: A utility perspective," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1025–1039, June 2015.

[14] 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, 2012.

[15] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1340–1351, 2013.

[16] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Joint optimization of user association and dynamic tdd for ultra-dense networks," in *IEEE INFOCOM 2018*, April 2018, pp. 2681–2689.

[17] G. Arvanitakis, T. Spyropoulos, and F. Kaltenberger, "An analytical model for flow-level performance in heterogeneous wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1488–1501, 2018.

[18] S. V. Hanly, C. Liu, and P. Whiting, "Capacity and stable scheduling in heterogeneous wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1266–1279, 2015.

[19] S. Borst, H. Bakker, M. Gruber, S. Klein, and P. Whiting, "Flow-level capacity and performance in hetnets," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*. IEEE, 2015, pp. 1–6.

[20] H. Kim, G. De Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, 2011.

[21] F. Wang, W. Chen, H. Tang, and Q. Wu, "Joint optimization of user association, subchannel allocation, and power allocation in multi-cell multi-association ofdma heterogeneous networks," *IEEE Transactions on Communications*, vol. 65, no. 06, pp. 2672–2684, 2017.

[22] L. P. Qian, Y. J. A. Zhang, Y. Wu, and J. Chen, "Joint base station association and power control via benders' decomposition," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1651–1665, April 2013.

[23] Y. Chen, J. Li, W. Chen, Z. Lin, and B. Vucetic, "Joint user association and resource allocation in the downlink of heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5701–5706, July 2016.

[24] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1040–1054, 2015.

[25] S. Borst, S. Hanly, and P. Whiting, "Optimal resource allocation in hetnets," in *2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 5437–5441.

[26] C. Liu, P. Whiting, and S. V. Hanly, "Joint resource allocation and user association in downlink three-tier heterogeneous networks," in *2014 IEEE Global Communications Conference*. IEEE, 2014, pp. 4232–4238.